PHYLOGENETICS AND HOMOLOGY MODELING

by

ALLEN WATKINS SMITH

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

and

University of Medicine and Dentistry of New Jersey,

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Microbiology and Molecular Genetics.

Written under the direction of

Peter Charles Kahn

And approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2008

ABSTRACT OF THE DISSERTATION

PHYLOGENETICS AND HOMOLOGY MODELING

By ALLEN WATKINS SMITH

Dissertation Director:

Peter Charles Kahn

Phylogenetics uses nucleotide and/or amino acid sequences to construct evolutionary trees and reconstruct the sequences (or other characteristics) of ancestral organisms. Proteins function almost entirely in their folded form, but phylogenetic work typically does not *directly* consider the structures into which protein sequences fold. Homology modeling uses a known protein structure to model the structure of a similar sequence, with the similarity arising from an evolutionary relationship - thus "homology". However, homology modeling typically does not *explicitly* use evolutionary data, even though the modeled proteins are part of evolved biological systems. Combining these fields is likely to be fruitful: since proteins are the product of organismal evolution, an examination of evolution is needed to understand them; since proteins are a vital component of all known organisms, an examination of protein evolution is needed to understand organismal evolution. Protein structure is more conserved than

protein sequence, especially for vital proteins. Therefore, the structure of a putative ancestral protein is likely to be close enough to modern-day structures to be modeled, especially if done in short evolutionary stages with each step having few sequence differences. It should therefore be possible to go down a tree, homology modeling the structure of a protein at each stage, then go back up again to a modern-day sequence to derive a structure for said sequence (usable as a test if already experimentally known). While the latter point has not been reached, considerable progress has been made. Ways in which structural data can assist in phylogenetics, such as whether predicted ancestral sequences are structurally realistic, have been found. A database of manually reviewed structural alignments of a variety of interesting proteins (with additional sequence alignments) has been created, as has a database of structures versus species. Some interesting phylogenetic findings have been made and a supertree construction technique explored. The phylogenetic program MrBayes has been enhanced, as have been the alignment capabilities of the program HMMer. An open-source suite of programs for homology modeling and phylogenetic analysis has been created; while not as automated as is desirable, these programs may serve as the basis for future work.

Acronyms and Abbreviations

We apologize for the usage of some non-standard abbreviations. These are due to either room considerations for tables (e.g., “Arith. M.” and “SA”) or desiring to make the PDF version of this dissertation more accessible to screen readers (e.g., not using special characters such as “Å” for Angstroms). The below table also contains some abbreviations that are standard in one field of this research but not others (e.g., “PDB” is a standard acronym in the field of biochemistry).

Acronym/Abbreviation	Meaning
3D	Three-Dimensional
ADH	Alcohol Dehydrogenase
ADH1	Alcohol Dehydrogenase Class I
Ang.	Angstroms (Å)
Arith. M.	Arithmetic Mean
CD	Circular Dichroism
C. albicans	*Candida albicans*
C. briggsae	*Caenorhabditis briggsae*
C. elegans	*Caenorhabditis elegans*
C. glabrata	*Candida glabrata*
deg.	degrees (º)
DHFR	Dihydrofolate Reductase
D. discoideum	*Dictyostelium discoideum*
eIF2a	Eukaryotic Initiation Factor 2a
eIF4e	Eukaryotic Initiation Factor 4e
eIF6	Eukaryotic Initiation Factor 6
eTF2a	Eukaryotic Termination Factor 2a
EC E.C.	Enzyme Commission Number (IUBMB 1992)
E. histolytica	*Entamoeba histolytica*
FFT	Fast Fourier Transform
G. gallus	*Gallus gallus* (chicken)

Acronym/Abbreviation	Meaning
GH	Glycosyl/Glycoside Hydrolase
GST	Glutathione-S-Transferase
GTR	General Time Reversible (transition matrix)
Harmon. M.	Harmonic Mean
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
MRCA	Most Recent Common Ancestor
NP	Non-Polynomial
ORO	Orotidine-5'-phosphate decarboxylase
P. carinii	*Pneumocystis carinii*
PDB	Protein Data Bank (Berman *et al.* 2000)
P. falcip.	*Plasmodium falciparum*
P. vivax	*Plasmodium vivax*
RMS	Root Mean Square
RMSD	Root Mean Square Deviation/Distance
S. cerevisiae	*Saccharomyces cerevisiae*
S. pombe	*Schizosaccharomyces pombe*
SA	Simulated Annealing
SOD	Superoxide Dismutase
TBP	TATA-Binding Protein (TF2D)
TPIS	Triosephosphate Isomerase
TS	Thymidylate Synthase
UBC	Ubiquitin Conjugating Enzyme
VdW	Van der Waals
vs.	versus

Dedication

To my spouse, Liora Engel (soon to be Liora Engel-Smith). Hopefully yours won't take as long, my love...

Table of Contents

List of Illustrations

The format for the figure titles has the Chapter number or Appendix letter as the first part; for figures starting with “4.T”, please see under “Tree results”, on page 202, for more information about the format.

Chapter 1: Introduction and Literature Review

1. Summary

Phylogenetics uses nucleotide and/or amino acid sequences of proteins to construct evolutionary trees and reconstruct the sequences (and/or other characteristics, e.g., behavior or morphology) in ancestral organisms (Nei, Zhang, & Yokoyama 1997). Proteins function almost entirely in their folded form, but phylogenetic work typically does not *directly* take into account the structures into which protein sequences fold. Homology modeling uses a known protein structure to model the structure of a similar sequence, with the similarity due to an evolutionary relationship[1] - thus "homology" (Eisenhaber, Persson, & Argos 1995; Marti-Renom *et al.* 2000). However, homology modeling typically does not *explicitly* use evolutionary data, despite that the proteins typically studied by it are part of biological systems, and, as Dobzhansky wrote, "Nothing in biology makes sense except in the light of evolution" (Dobzhansky 1973). Combining these fields is likely to be fruitful: since the proteins most of interest are the product of evolved biological systems, an examination of protein evolution is needed to understand them; since proteins are a vital component of all known organisms, an examination of protein evolution is necessary to examine fully organismal evolution.

[1] See footnote 27 under "4. Connecting Phylogenetics and Homology Modeling: Critical Questions", on page 16, for more discussion of why, when modeling is successful, an evolutionary relationship is highly likely.

Protein structure is more conserved than protein sequence, especially for vital proteins (Rossmann, Moras, & Olsen 1974). Therefore, the structure of a putative ancestral protein is likely to be close enough to modern-day structures to be modeled, especially if done in stages with each evolutionary step having few sequence differences. It should therefore be possible to go down a tree, homology modeling the structure of a protein at each stage, then go back up again to a modern-day sequence to derive a modeled structure for the modern sequence. This model would be usable as a test of the entire process if the 3D structure of the modern sequence were already experimentally known.

2. Phylogenetics - Ancestral Sequence Prediction

In phylogenetics, one possibility is to determine a probable ancestral sequence and then examine it for properties of interest (Nei, Zhang, & Yokoyama 1997). Some efforts at relating ancestral protein differences to structural changes have been made. Until very recently, such efforts have been primarily or entirely through either:

- The examination of the location of amino acid changes in modern-day protein examples (Chandrasekharan *et al.* 1996; Dean & Golding 1997; Miyazaki *et al.* 2001); or
- X-ray or NMR examination of moderately - i.e., without all the changes needed to reach the predicted ancestral state - mutated modern proteins (Hurley, Chen, & Dean 1996; Wilson, Malcolm, & Matthews 1992).

The exception (Ortlund *et al.* 2007) took place after the present research, including the origination of all (to our knowledge) duplicated ideas, was well under way (Smith & Kahn 2005); the other research also used a comparatively evolutionarily recent hypothetical ancestral sequence for its structural work.

As compared with those scientists who study organisms on the morphological level, we have had the disadvantage that, with a few limited[2] and problematic[3] and/or evolutionarily recent[4] exceptions (Asara *et al.* 2007; Wayne, Leonard, & Cooper 1999), no remains of ancient molecules survive, unlike fossils. Essentially all that we have had to study are the present-day results of the evolutionary process. Given this, it is perhaps unsurprising that many current scientific debates over evolution are over molecular evolution, such as the degree to which (apparently) neutral mutations - which are especially likely to be detectable only on the molecular level - play a role (Kimura 1983). (For an example of a tree with present-day and (predicted) ancestral sequences - taken from the present work[5] - please see Figure 1.1, on page 4.)

---

[2] For instance, while *Tyrannosaurus rex* protein sequences have recently been determined (Asara *et al.* 2007), they are limited to proteins found in high concentration in bones (e.g., collagen).

[3] Among the problematic issues are those of potential contamination (Walden & Robertson 1997).

[4] At least, evolutionarily recent in comparison to the present research.

[5] The protein shown is residues 46-59 (as per the alignment in “Appendix K: Partial DHFR alignment”, on page 384) of DHFR.

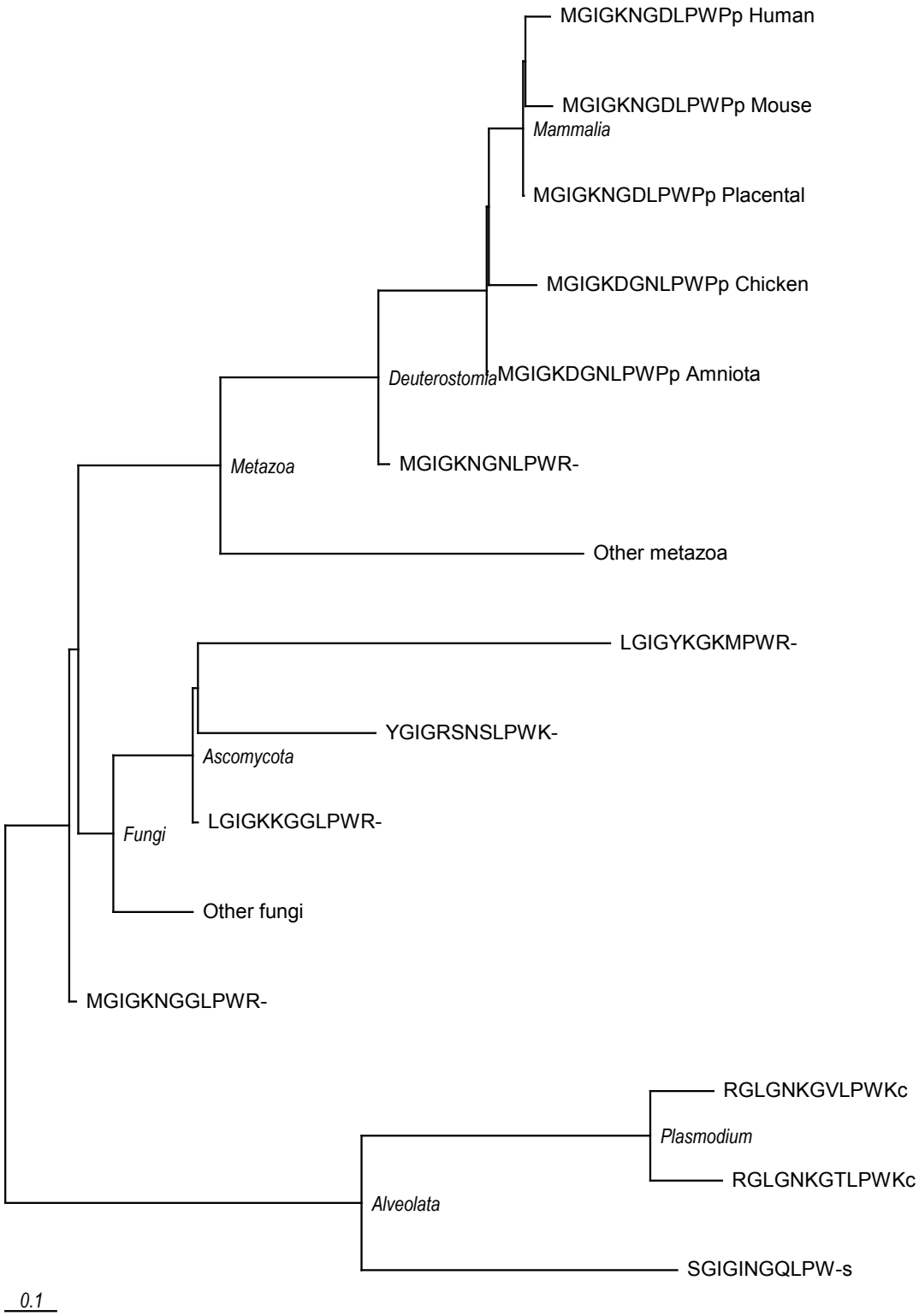


Figure 1.1: Example tree with predicted ancestral sequences

Unfortunately, the reconstruction of an ancestral sequence has a number of possible pitfalls, generally centering on the problem of determining the likely mutations taking place in the sequence of a protein. This information is necessary both for the determination of a phylogenetic tree and for the prediction of the sequence at the branch points in that tree. In general[6], the methods currently in use do not allow for differences in likely mutations due to the surrounding amino acids[7]. Some methods in use fail to allow for any variation in likelihood of mutations at all, even when the variation is independent of nearby amino acids.

Moreover, such predictions are generally based on mutational likelihood information derived from sequence alignments. In turn, the alignments are based on previously gathered information on mutational likelihood, on manual (visual) alignment, or on structural alignment. Manual alignment is subjective, time-consuming, error-prone, and assumes that the person doing the alignment knows exactly what is of significance in a protein's sequence and what is not. Finding an optimal structural alignment is at least an NP-complete problem, and may be an NP-hard problem (de la Higuera & Casacuberta 2000; Lathrop 1994; Lathrop *et al.* 1998; Westhead *et al.* 1995).[8] This (apparently theoretical) consideration is

[6] The exceptions mainly (Fornasari, Parisi, & Echave 2002) involve examination and classification of the overall surrounding environment - e.g., level of hydrophobicity, type of structure, or degree of surface exposure (Goldman, Thorne, & Jones 1998; Koshi & Goldstein 1995; Overington *et al.* 1990; Overington *et al.* 1992; Robinson, D M *et al.* 2003; Wako & Blundell 1994a, 1994b).

[7] Also not taken into account by *most* methods are the current codons for the amino acids in question, although they typically take into account that some codons are easier to mutate into others based on the codons themselves alone. Matrices like these, e.g., BLOSUM62 (Henikoff & Henikoff 1992), for amino acids alone, also do not take into account genetic code changes (Massey *et al.* 2003; Telford *et al.* 2000).

[8] NP-complete means that, with all currently known methods of solving the problem, the time

borne out by the fact that multiple equally preferable structural alignments - with very different sequence alignments - can be found in many cases (Godzik 1996). Moreover, criteria for structural alignment can be more arbitrary than they would appear at first glance (Falicov & Cohen 1996; Gerstein & Levitt 1996; Levitt & Gerstein 1998; Yang, A-S & Honig 2000a, 2000b, 2000c; Zemla *et al.* 1997).

Despite these difficulties, structural alignment with manual inspection (as is used in this work - see “3b. Alignment of other sequences”, on page 78) is generally considered the best method of alignment, and indeed has been used to judge the quality of alignments from other methods (Domingues *et al.* 2000; Gerstein & Levitt 1998; Jaroszewski, Rychlewski, & Godzik 2000; Sauder, Arthur, & Dunbrack 2000). It is, however, limited in its application to sequences with known

---

needed to solve the problem goes up faster than a polynomial (P) (e.g., x, x-squared, x-cubed) with increasing size of the problem (e.g., the size of the proteins to be aligned, measured in arbitrary units). (If the time were to go up at only a polynomial rate, then this would be considered solvable in "polynomial time".) For instance, if a problem involving an amino acid chain took 100 seconds for 10 amino acids, or 400 seconds for 20 amino acids, it appears to be of polynomial complexity (i.e., can be solved in polynomial time) - the polynomial happens to be x-squared in this case. The time required to solve an NP-complete problem goes up at a rate faster than this, or any higher-order polynomial. For instance, if an alignment problem took 1024 seconds for 10 amino acids, but 1048576 seconds for 20 amino acids - a rate of 2 to the x - and no faster solution was reliably locatable now, the problem would be NP-complete. (A similar comparison is between exponential (e.g., doubling with each generation) and linear (e.g., increasing by 10 units with each generation) growth, in which - over sufficient time - exponential growth will always outpace linear growth, no matter how low the base of the exponential growth. An example of this problem is unrestricted population growth compared to productive capacity for food (Malthus 1798).) Finding a means to solve an NP-complete problem in polynomial time will solve all other NP-complete problems in polynomial time (any NP-complete problem solution can be transformed into a solution for all other NP-complete problems). Given the amount of work that has gone into this without much success, it appears unlikely that this will be done anytime soon. An NP-hard problem is one that not only is not solvable in polynomial time by current methods, but a solution to NP-complete problems in polynomial time will not solve it (although finding a polynomial-time solution to an NP-hard problem will solve all NP-complete problems in polynomial time) (Lopez-Ortiz 2000). NP-complete problems are usually handled either heuristically (using methods that are not guaranteed to find the best solution - "best guesses"), via "brute-force" searches through all the possibilities, or via combinations of these (e.g., using heuristics to eliminate some possibilities from the brute-force search).

structures. Some attempts (threading[9]) have been made to align sequences with unknown structures to sequences with known structures making use of the structural information. Unfortunately, most threading methods appear to be better at recognizing folds (i.e., recognizing that a sequence is likely to fold into a structure similar to a particular known one) than at generating good alignments (Bienkowska *et al.* 2000; Sunyaev *et al.* 1998); moreover, the protein threading problem is itself NP-complete in difficulty (Lathrop 1994).

In general, the likelihood of a particular mutation happening is expressed by a matrix; such a matrix may be of bases, amino acids, codon triplets, or higher structural features (Cootes *et al.* 1998; Eck & Dayhoff 1966; Henikoff & Henikoff 1992, 2000; Koshi & Goldstein 1995; Overington *et al.* 1992; Wako & Blundell 1994a, 1994b; Yang, Z, Nielsen, & Hasegawa 1998). Sequence alignment can be defined as finding the way of putting two or more sequences next to each other so that the likelihood of the evolutionary transitions between them is maximized. In other words, the residues that an alignment shows as corresponding to each other in two sequences are, ideally, residues that have a common evolutionary origin - they descended from a single common ancestral

[9] In threading, a protein sequence is "threaded" through a known protein structure, and the compatibility between the sequence and the structure is tested using various scoring schemes (e.g., hydrophobic residues should not be on the surface of the protein). (It is called "threading" because one can look at it as if the existing residues in the structure were a tube and the new sequence was a thread being passed through the tube.) This procedure is then - at least for threading used for fold recognition - repeated with other structures, and which structure is most compatible with the sequence is determined. (Sunyaev *et al.* 1997; Wikipedia 2006; Zhang, C & Kim 2000)

position, i.e., are homologous. The likelihood of the individual transitions is estimated using one or (rarely) more[10] matrices.

One type of transition that can be allowed in a matrix, or more usually separately, is the introduction or extension of a gap[11]. The methods of, and parameters for the methods of, allowing for gaps are currently one of the more arbitrary areas of sequence alignment, especially when used for global alignment (Abagyan & Batalov 1997; Golubchik *et al.* 2007; Henikoff & Henikoff 2000; Kjer 1995; Vogt, Etzold, & Argos 1995); manual editing of automatically-determined gaps (using functional and/or structural information) is frequently necessary.

The construction of phylogenetic trees is also known to have errors from various sources, including variable rates of mutations, homoplasy (such as convergent[12] and parallel evolution), and diversity inside species (Brower, DeSalle, & Vogler 1996; Lanyon 1993; Philippe & Laurent 1998; Yang, Z 1996a). If the tree on which an ancestral sequence reconstruction is based is incorrect, then the ancestral sequence reconstruction is likely to be incorrect (Ronquist 2004). At

[10] More than one matrix (or the use of a weighted mixture of matrices) can be used in the case of, for instance, allowing for the effects of different secondary structure types (Lartillot, Brinkmann, & Philippe 2007; Lio & Goldman 1998; Lio *et al.* 1998; Overington *et al.* 1992), or when using data from both protein and DNA at once (Arvestad 1997, 1999).

[11] Gaps can occur when a base pair insertion or deletion has taken place, generally in non-coding DNA, when a mutation has altered intron splicing (in a eukaryote), or a recombination event has removed or duplicated some bases. (Base pair insertions or deletions in coding DNA will induce a frame shift, unless they add up to a multiple of three; changes of 1 or 2 base pairs are selected against in coding DNA by the "nonsense" protein section produced. This section can be either until the end of the protein or just until more insertions or deletions add up to a multiple of three.)

[12] Convergent evolution happens when unrelated proteins evolve to become more similar, generally due to functional constraints (e.g., enzymatic activity or interfacing with another protein). Parallel evolution happens when the same mutations happen independently in two species, again generally due to functional constraints (e.g., common environmental changes); this can appear to be due to the two species diverging later than was actually the case (Futuyma 1986).

best, it will be a reconstruction of what that sequence would have (probably) been if the organisms *had* evolved according to that tree. However, the error is likely to be significant only if the tree is in error in a region close by (e.g., is descended from) the ancestral node of interest (Zhang, J & Nei 1997).

Another potential source of error is the matrix (of amino acid or nucleotide replacement likelihoods) used in constructing a tree. Matrices are used in the construction of most types of phylogenetic trees[13] as well as in the sequence alignments that are necessary before the construction of the tree. Trees are constructed to maximize the likelihood of the transitions[14] taking place from (hypothesized) common ancestral sequences to the known sequences (in extant species) (Brower, DeSalle, & Vogler 1996; Cavalli-Sforza & Edwards 1967; Edwards, A W F & Cavalli-Sforza 1964; Farris 1977, 1983; Felsenstein 1984a, 1984b; Fitch & Margoliash 1967; Higgins 2000; Thornton & DeSalle 2000). For studies looking at a very wide range of time scales, matrices combining DNA and protein (Arvestad 1997, 1999), using the information from DNA of what base pair and other changes are likely[15] and from protein evolution of what amino acid

[13] Parsimony trees do not use matrices except for the initial alignment; they usually only look at minimizing the number of mutations, not the likelihood of said mutations. The major exception is weighted parsimony (Felsenstein 1981).

[14] Another way to look at this is that their construction process attempts to maximize the inverse correlation between the likelihood of a particular set of changes "between" (or, more precisely, from a common ancestral state to) two points and the distance on the tree between those two points.

[15] For instance:

- Transitions (purine-to-purine or pyrimidine-to-pyrimidine) and transversions (purine to pyrimidine or vice-versa) typically happen - or at least are evolutionarily accepted - at different rates (Keller, Benasasson, & Nichols 2007; Sommer 1992; Zhang, Z & Gerstein 2003). This pattern is particularly found in protein-coding genes due to the biases in the genetic code regarding which mutations are synonymous (Huelsenbeck & Nielsen 1999). Another factor is that transitions are promoted by methylation at CpG sites (Keller, Benasasson, & Nichols 2007).

substitutions are likely, are particularly desirable. This desirability is because some information is lost if one does not use both. If protein alone is used, for instance, most information regarding third base pairs in DNA is lost. If DNA alone is used, the likelihood of changes from one amino acid to another is lost. Such information is particularly of importance over short evolutionary distances, since DNA changes faster than protein[16] and thus is better at tracking fast changes, but will be overwhelmed by noise for slower changes (Goldman & Yang 1994; Kreitman & Comeron 1999; Muse & Gaut 1994; Yang, Z *et al.* 2000).

Given the above problems, it is known that the reconstructed ancestral sequences are likely to have a number of errors from a variety of sources (Cunningham, Omland, & Oakley 1998; Zhang, J & Nei 1997), particularly if using parsimony (Collins, Wimberger, & Naylor 1994). One way of estimating the likelihood of errors[17] in a phylogeny is by deleting some information (e.g., some bases or amino acids in a sequence), possibly replacing the information with other randomly selected positions, and seeing if the constructed tree remains the same[18]. This method is known as bootstrapping - or, without replacement, jack-

---

- Different codons can require more or fewer changes to go from one amino acid to another. Even DNA mutations that do not change the amino acid sequence may make a later change easier.

These biases are, as noted above, less likely to make a difference at longer time scales (Brown, W M *et al.* 1982).

[16] For protein coding sequences, this difference is likely to be due to the redundancy of the genetic code - while codon preferences may cause some constraints, a change from one codon to another codon in which both codons code for the same amino acid has rather less effect than a change resulting in different amino acids.

[17] To be more precise, bootstrapping is a way to estimate the support for the phylogeny. It can, however, be considered a means of detecting the error of relying too much on a small portion of the available information.

[18] Areas in a tree that change when information is deleted are considered less reliable - there is less support for that area and, given this lack of support, often a greater likelihood of error in that area of the tree, since it is from less information.

knifing (Efron 1979; Felsenstein 1985a, 1988; Lake 1995). Bootstrapping can help estimate the likelihood of errors due to oddities in the original sequence and due to sampling error in the construction of the matrices involved. However, bootstrapping is either unlikely or unable (depending on the source of error) to help in estimating the likelihood of error due to other factors (Cummings, Otto, & Wakeley 1995; Kunsch 1989; Peng *et al.* 1992; Sanderson 1995). For instance, errors due to correlations between different positions in the sequence will only be detected if one happens to delete both correlating positions (Chang, B S W & Campbell 2000; Galtier 2004). Moreover, bootstrapping is computationally infeasible with many methods of phylogenetic tree determination, including those used in the present study.

One method of decreasing the errors in the construction of a phylogeny is by using more than one protein (or other sequences, such as rRNA) in constructing it (Bull *et al.* 1993; Cummings, Otto, & Wakeley 1995, 1999; Otto, Cummings, & Wakeley 1996; Russo, Takezaki, & Nei 1996). Phylogenetic trees are constructed using a model of what evolutionary occurrences are most likely (e.g., a minimal number of changes for the parsimony model) that is used as a criterion to decide which tree (or set of possible trees) is most likely. As well as the possibility of this model being incorrect in general (covered above), there is also the possibility that it is incorrect for a particular set of sequences[19]. By using

[19] For instance, a particular protein's gene may be the subject of horizontal gene transfer with respect to (most of) the other genes in the species, meaning that for that protein/gene there would be a difference between the gene tree and the species tree. Admittedly, if horizontal gene transfer is sufficiently frequent, the “species” of the “species tree” are uncertain (Gogarten & Townsend 2005).

more than one group of homologous sequences, one reduces the chance of this variety of error.[20]

The usage of a known 3D structure for one or more modern-day variants of a protein may assist in determining the likelihood of various putative ancestral sequences, and eventually in the process of producing both:

- the phylogenetic trees on which those sequence predictions are based; and
- the predicted sequences themselves.

For instance, one major cause for inaccuracies may be a failure to allow for variations in likelihood of mutations due to the effects of surrounding residues. In this, "surrounding" does not only include those residues close in the sequence, but those that are close in the 3D structure (Cootes *et al.* 1998; Dutheil & Galtier 2007; Fukami-Kobayashi, Schreiber, & Benner 2002; Gaucher, Miyamoto, & Benner 2001; Gobel *et al.* 1994; Golding & Dean 1998; Peng *et al.* 1992; Pollock, Taylor, & Goldman 1999; Saraf, Moore, & Maranas 2003; Singer, Vriend, & Bywater 2002; Wilson, Malcolm, & Matthews 1992). These correlations, if present, are also of interest with regard to bootstrapping, as noted on page 11.

---

[20] Bootstrapping can help estimate the likelihood of the method being incorrect for a particular sequence, *if* the errors are due to part of the sequence only (e.g., if the error is that the method is relying too much on that part of the sequence, and that part gives results different from other areas of the sequence). If bootstrapping indicates that a tree constructed using only one sequence source (e.g., only one protein, found in multiple species) may be in error, then an expansion of the data to encompass more than one protein (sequence source) is recommended (Efron 1979; Felsenstein 1985a; Kunsch 1989; Lake 1995). Of course, it is possible that part of the sequences used is correct in terms of the evolutionary history of the species but the majority is incorrect. This depends, however, on:

- how representative the sequences used are of the entire genome of the species; and
- how one defines species.

For instance, if the sequences are representative of the genome of the species, and one defines a species by its genome (as is implicit in the usual definition of a species as a reproductively isolated set of organisms), then the evolutionary path of most of the sequences is the

3. Homology Modeling

Homology modeling is the process by which one or more proteins with known structures, with sequences similar to a protein of interest that lacks a known structure, are used to model the unknown structure (Eisenhaber, Persson, & Argos 1995; Goldsmith-Fischman & Honig 2003; Lipke *et al.* 1995; Sanchez & Sali 1997a). Methods of modeling protein structure are needed because we have far more sequences available than we have structures. Moreover, the ratio between the number of sequences known and the number of structures known is getting greater all the time, as is its rate of increase (Bowie, Luthy, & Eisenberg 1991; Goldsmith-Fischman & Honig 2003; Mosimann, Meleshko, & James 1995; Rost & Sander 1996)[21]. Modeling structures based only on a sequence is a NP-complete problem (Berger & Leighton 1998; Crescenzi *et al.* 1998); it is (for a reasonable degree of quality[22]) computationally infeasible for all but the shortest single[23] sequences (Bonneau & Baker 2001; Defay & Cohen 1995).

---

evolutionary path of the species.

[21] Some evidence suggests that there are only a limited number of protein folding patterns ("folds") found in nature (D'Alfonso, Tramontano, & Lahm 2001; Overington *et al.* 1990). If at least one example of each fold were to be structurally determined, then it would theoretically be possible to use this data and "homology" modeling to determine the structures of all other proteins. Such a possibility would be dependent on either:

A. being able to recognize (e.g., via threading - see footnote 9 under "2. Phylogenetics - Ancestral Sequence Prediction", on page 7) what known fold new sequences would fold into, then using the techniques of homology modeling without necessarily having an evolutionary relationship present; or, perhaps more likely,
B. having in a database at least one homologous protein with a known structure - something that should be doable given the common ancestry of all known living things - and being able to recognize the homology in question. This recognition would be easiest to do if a common function is known and/or the sequence similarity is high enough to make a structure with a different fold unlikely.

This goal is one motivation for "structural genomics" - getting the structures for a wide variety of proteins found in the genomes of many organisms (Goldsmith-Fischman & Honig 2003).

[22] By a "reasonable degree of quality" is meant a backbone alignment versus experimentally determined structures for the same sequence with a moderate-to-low RMSD - e.g., significantly

Homology modeling is used to get around this problem, but is not generally possible below 20% sequence identity, is extremely difficult below 40% identity, and even at somewhat higher sequence identities is likely to be inaccurate (Bowie, Luthy, & Eisenberg 1991; Chung & Subbiah 1996; Rost 1999; Sternberg *et al.* 1999; Taylor 1994; Taylor, Flores, & Orengo 1994)[24]. Automated modeling procedures, despite their considerable advantages in terms of time and reduction of human labor, are even more dependent[25] on a high level of sequence identity to be accurate (Bowie, Luthy, & Eisenberg 1991; Dalton & Jackson 2007; Mosimann, Meleshko, & James 1995; Sanchez & Sali 1997b; Saqi, Russell, & Sternberg 1998; Taylor 1994; Winn *et al.* 2004).

In homology modeling, amino acids in a known structure (the "template) are substituted with those in a sequence of unknown structure (or amino acids not present in the sequence of unknown structure are deleted). If sections of the sequence of interest are not found in the template, these are inserted from other

---

below 3.626 Ang., the level expected for two structures with only a chance 5% of residues identical (Vogt, Etzold, & Argos 1995).

[23] The use of evolutionary information (including as implied in sequence alignments) can assist in "*ab initio*"/"*de novo*" structural prediction (Ortiz *et al.* 1999).

[24] One reason for this is likely to be the significant dependence of local conformations on the global structure of the protein - some identical sequences (of significant length) are found to adopt markedly different configurations (alpha-helical versus beta-sheet) in different structures (Jacoboni *et al.* 2000; Zhao *et al.* 2001). Moreover, as shown in the Paracelsus challenge,, even a change of 50% or less of the residues (a sequence identity of 50 %+) is capable of transforming a protein between all beta and all alpha (Dalal, Balasubramanian, & Regan 1997; Rose & Creamer 1994).. Knowing the general "fold" of the protein (see footnote 21, on page 13) may be of assistance by telling something about the global structure - whether it is sufficient to overcome this problem may vary.

[25] One reason for this limitation is the need for an accurate alignment. During the process of manual model building, it is more likely (as found with the present work; see "5. Alignment of central sequences", on page 336) that the human modeler will recognize an alignment problem (Dalton & Jackson 2007). Iterative automatic alignment and modeling is a potential alternative

(structurally known) proteins (a "loop search"). Conformational changes are then made to minimize the predicted potential energy of the structure (a many-body problem) and otherwise make the characteristics of the structure resemble those of native folded proteins in general[26].

4. Connecting Phylogenetics and Homology Modeling: Critical Questions

1. Starting with one or more known (modern) 3D structure(s), can we follow the tree of putative ancestral sequences backward (down one or several branches) and forward (on other branches), reconstructing the 3D structures of these sequences via homology modeling, and reach a correct modern-day structure?
2. If this is not the case, why not? If it is only sometimes the case - if it sometimes works and sometimes does not work - why? For instance, are some methods more reliable? Among the methods involved, how can we use a 3D model of the structure of a related sequence (e.g., one descended from it) to help predict - and/or estimate the validity of - an ancestral sequence?

These are examined in further detail below.

---

means of solving this problem (John & Sali 2003), but is computationally quite demanding and requires the ability to recognize bad models on an automated basis.

[26] The latter includes optimizing characteristics that are not additive in nature. This factor makes techniques such as Dead End Elimination much less useful (Betancourt & Thirumalai 2002; Clark & Westhead 1996; Desjarlais & Handel 1995; Desjarlais & Clarke 1998; Desmet, Spriet, & Lasters 2002; Hayes *et al.* 2002; Hinds & Levitt 1996; Kono & Saven 2001; Lazar, Desjarlais, & Handel 1997; Looger & Hellinga 2001; Tuffery, Etchebest, & Hazout 1997; Voigt, Gordon, & Mayo 2000; Voigt *et al.* 2001; Zou & Saven 2000, 2003).

The supposition at the heart of homology modeling is that proteins that are similar in sequence will be similar in structure (Eisenhaber, Persson, & Argos 1995; Goldsmith-Fischman & Honig 2003; Lipke *et al.* 1995; Sanchez & Sali 1997a). Why are proteins sometimes similar in sequence, even outside of the active site and other highly functionally constrained regions? Either the resemblance is by chance, or the proteins have a common ancestor. In the first case, as one would expect from its name, homology modeling is much less likely to be successful (Preisner *et al.* 1997; Reardon & Farber 1995; Russell *et al.* 1998; Saqi, Russell, & Sternberg 1998)[27]. If the proteins actually are homologous (the second case), and have a similar (or the same) function, then they are likely to retain a considerable degree of structure in common. Except when a change of function has taken place, structure is more conserved than sequence in

[27] While some successes have allegedly (depending on one's definition of success) been seen in modeling the overall structure of a protein from substituting a sequence into a known structure formed by a supposedly unrelated sequence (this can be considered an extension of a "loop search", the technique of substituting small fragments (generally of loop (so-called "random coil") regions) of another structure into otherwise-undetermined locations in a model (Deane & Blundell 2001; Zhang, Y P, Kolinski, & Skolnick 2003)), this approach:

1. Is limited to gross structural features and not to fine details or active site chemistry, unless experimental data are available regarding, e.g., the active site configuration (Gilquin *et al.* 2002);
2. May well work due to the limited number of possible gross protein folds (D'Alfonso, Tramontano, & Lahm 2001; Overington *et al.* 1990);
3. Leaves open the possibility that the two sequences in question are, in actuality, distantly homologous; and
4. Cannot work in all cases given the importance of tertiary structure (see footnote 24 under "3. Homology Modeling", on page 14).

Fortunately, resemblance by chance becomes less and less likely the longer the sequences in question. Convergent evolution and other forms of homoplasy (Futuyma 1986) are unlikely for proteins of significant size (Brower, DeSalle, & Vogler 1996; Rossmann, Moras, & Olsen 1974), at least for the protein as a whole (as opposed to, e.g., active site residues or residues interacting with other proteins). One way of looking at this is by means of Dollo's Law: complex characteristics, once lost, are unlikely to re-evolve in their original form (Dollo 1893; Farris 1977; Futuyma 1986). While gene sequences are not sufficiently complex for this to be the case in general – reversals must be allowed for in sequence evolution (Felsenstein 1984a; Kimura 1983) – structures are another matter in many cases.

evolution (Rossmann, Moras, & Olsen 1974)[28]. Therefore, the structure of an ancestor of a protein should be similar to the present-day structure of that protein. This similarity implies that if we can accurately deduce the ancestral sequence, and know the present-day structure of the protein, it should be possible to model the ancestral structure. Moreover, the reverse (going from a modeled ancestral structure to the structure of a present-day sequence) should likewise be possible.

Of interest are ways to use an experimentally-determined structure - or an already-constructed (homology) model of the structure - of a related sequence to help predict an ancestral sequence, and to estimate the likelihood of other predictions being correct.[29] As well as reducing the computational effort of the modeling involved, this usage is of importance in selecting which theorized ancestral proteins one should investigate further (e.g., via paleomolecular biochemistry - see "Paleomolecular biochemistry", on page 364). Another way in which structures may be used is in helping to refine the alignment, such as by examining which residues appear to be performing which function (e.g., ligand binding).

---

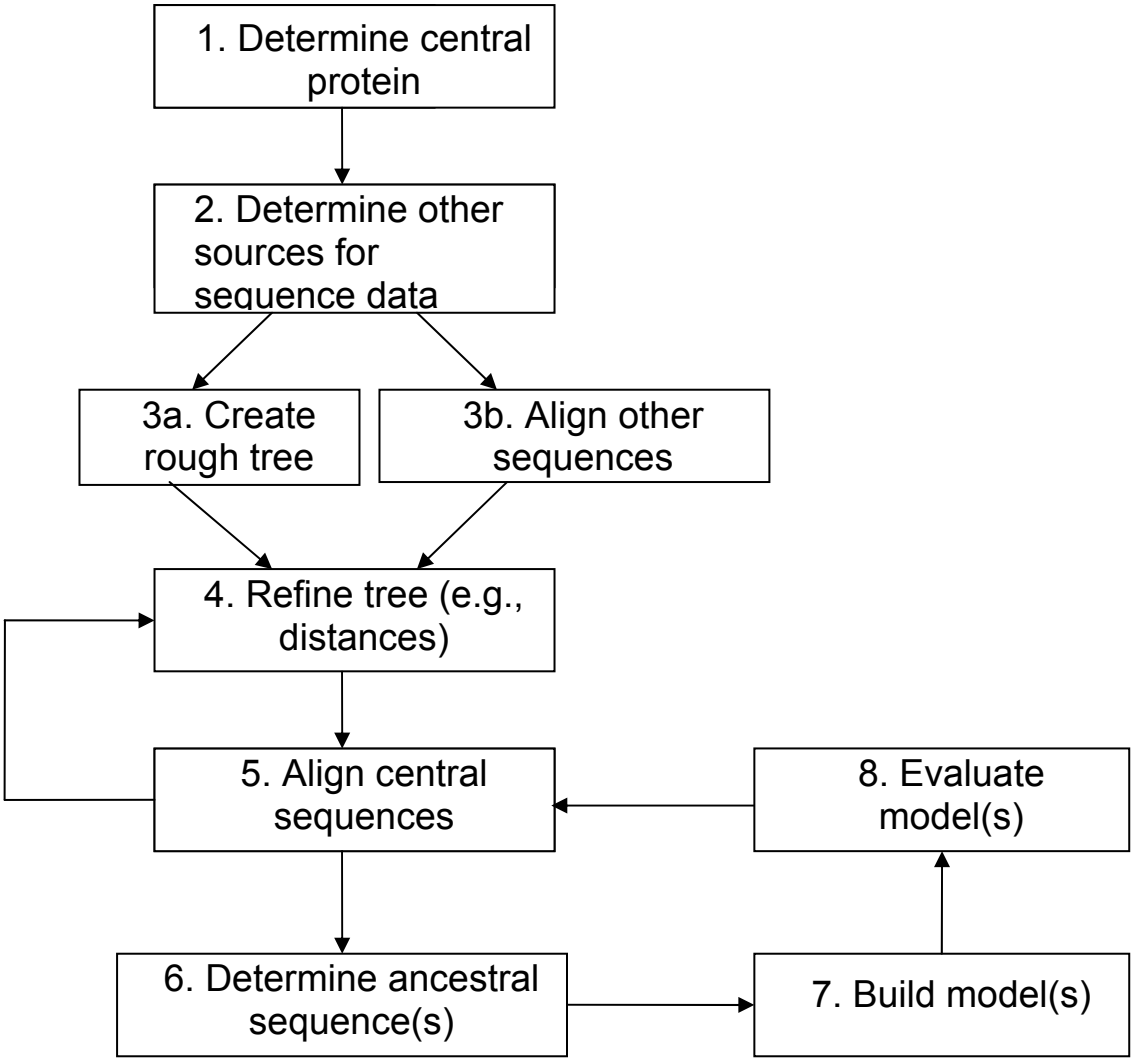
[28] However, this may only hold true for structures undergoing active evolution. Pseudogenes may keep their original form for quite some time (Marshall, Raff, & Raff 1994); reactivated pseudogenes may produce a new protein with as much divergence in structure as in sequence from the ancestral protein, particularly if the protein's function shifts.

[29] I.e., we have attempted to improve the phylogenetic prediction of one sequence by looking at the 3D structure of a related sequence. For instance, we have examined places where an amino acid predicted as a possibility by phylogenetics will not function in the modeled structure, thus narrowing down the possible amino acids at a location.

Chapter 2: Research Design

The flowchart shown in Figure 2.1, below, is a summary of the research design. Each item in it is explained and expanded upon further below (e.g., definitions of "central protein" and "other sources/sequences"); most sections in this dissertation are named (and numbered) correspondingly.

Figure 2.1: Research Design



1. Determination of central protein

We need to designate a "central" protein for initial analysis (i.e., for the test of the ancestral sequence and homology modeling procedure). There are several needed characteristics for the protein in question:

A. At least at the start, it should be a single-domain protein. This criterion will help in keeping the size down - in general, single-domain proteins will be under 300 amino acids (Richardson, J S 1981, 2004-2006)[30].

B. Ideally, the protein should be monomeric, at least for the source and target species. This criterion and the previous one help in simplifying the process by removing protein-protein interactions.[31]

C. The protein should not be of a class that is little understood and so hard to homology model even for close-by cases. One example of such a class is that of membrane proteins; these should be avoided.

D. Similarly, the protein should be easily producible and characterizable in the laboratory[32]; for instance, a protein without significant repetitive secondary structure would not be suitable, since it would be difficult to get information about its structure from CD[33] or other methods. Like the prior

[30] This number does depend on one's definition of domains.

[31] This criterion will probably be one of the first to disappear (it is present for the current work because this project is a proof of concept, with a wish to avoid unnecessary difficulties). Relatively few proteins in the PDB (i.e., proteins whose structures are publicly available) are all of single-domain, of small size, functionally vital, and with structures from multiple sources; many are from the same source with different ligands, at different resolutions, *etc.* Homodimers should be acceptable, since their interfaces can be simulated via mirroring.

[32] This criterion is for the sake of future work with the predicted ancestral sequences/structures.

[33] Circular Dichroism (CD) can indicate secondary structure proportions without as many difficulties as X-ray crystallography or NMR structural determination (although it does not reveal the full structure, unfortunately). The exact definition of secondary structures in full 3D structures that corresponds to that effectively used in the interpretation of CD data is under some dispute (Drennan 2001).

criterion, this one will tend to rule out membrane proteins (while their helical content is high, they are not easy to work with in the laboratory).

E. It should have homologues with known sequences (ideally protein sequences, to reduce sequencing error) from many different organisms, so that accurate ancestral sequences can be deduced.[34] Moreover, many other sequences that are usable for tree construction should be known from the organisms in question. Ideally, the organisms' full genomes should have been sequenced, and a number of protein structures (to use in alignment - see "Structural and initial sequence alignments" on page 128 - and to correct for sequencing errors) should be known from those organisms. Particularly of importance are sequences (and structures) from the organisms with the source and target structures.

F. We need sequences - and, preferably, structures - of the target protein from outside the clade (group of species descended from a common ancestor) containing the end-points of interest (i.e., outgroup sequences[35] - see “Appendix O: Outgroup review/explanation”, on page 412).

G. Either:

    1. It should be found in one copy only in each organism, or

[34] For future work, this criterion includes that we should have sufficient different sequence sources (i.e., species having the protein) that we are able to construct putative ancestral sequences using only a subset of the sequences. Such subsets are needed in order to make sure that findings of different accuracies of ancestral sequence reconstruction between different methods are not specific to a particular set of species (and are instead due to differing methods). (This idea can be regarded as a variety of bootstrapping or jack-knifing.)

[35] For future work, it may be possible to use previously diverged proteins or pseudogenes for this, if one can be adequately sure that they were indeed previously diverged. (However, there would be a worry as to whether they were too far diverged, even if alignment was possible (for a non-pseudogene) via structural information - see “Appendix O: Outgroup review/explanation”, on page 412, for some discussion of this problem.)

2. the different copies (isozymes) need to be different enough from one another that one can clearly tell at least one (target) group of homologues (one per species) that has descended more directly from a common ancestral sequence (from outside the area of the tree of interest) than have other proteins (Arvestad *et al.* 2003). For instance, myoglobin and hemoglobin from most organisms can be told apart easily, as can the various families of cellulases and xylanases (Coutinho, P M & Henrissat 1999; Coutinho, Pedro M & Henrissat 2007). However, glycosyl hydrolases of family 18 (chitinases) have both:
    a. subclasses that can be difficult to distinguish between (they are close enough together in plants to do homology modeling between subclasses) but have structural (disulfide presence/absence) and functional differences not readily apparent from the sequences (Parise 2005); and
    b. multiple known chitinases in several organisms[36] of interest.

H. The protein should not be highly polymorphic, particularly in source and target organisms. Sequences that are highly variable within species, such as HLA (Human Leukocyte Antigen) type genes (Gaur *et al.* 1992), may be problematic. Particularly to be avoided are ones with as much divergence within individual well-defined[37] species as between closely

[36] E.g., *Aspergillus fumigatus*, according to a `blastp` search - see “Structures and sequences”, on page 61 - followed by manual examination (data not shown).

[37] In "well-defined" species, organisms reproduce with other organisms inside the species but not with organisms outside the species. Bacteria are examples of organisms not qualifying on either ground (Gogarten & Townsend 2005); many plants do not qualify on the ground of reproduction

related species (Brower, DeSalle, & Vogler 1996; Bull *et al.* 1993; Page 1993).

I. It needs to be one that has kept common functions and structural elements for a wide span of evolutionary divergence. Moreover, it needs to be a necessary component for the life of the organism, so that it is reasonable to assume that its ancestors were likewise similar.[38]

J. The protein should not have been the subject of horizontal gene transfer. Examples of horizontally transferred proteins include aerobic metabolism enzymes (which may have transferred from mitochondria) and syncytin (involved in human placental morphogenesis) from a captive retrovirus (Bensasson, Zhang, & Hewitt 2000; Berg & Kurland 2000; Mi *et al.* 2000; Shafer *et al.* 1999; Spolsky & Uzzell 1984; Takahata & Slatkin 1984).[39] These may not be suitable, since their ancestral trees will diverge from that for other sequences at the point of incorporation (Brower, DeSalle, & Vogler 1996; Bull *et al.* 1993; Gogarten, Doolittle, & Lawrence 2002; Gogarten & Townsend 2005; von Haeseler & Churchill 1993; Nelson 1983; Page 1993; Wanntorp 1983; Xu 2000)[40]. However, this problem may be simply avoided by not going back that far. For instance, if using a

---

frequently occurring outside the species; non/seldom-sexual fungi do not qualify on the ground of insufficient reproduction within the species to be sure of the species definition.

[38] For future research, enzymes involved in glycolysis may be suitable, except for the problem that most of them are multi-domain.

[39] Even more nonfunctional DNA - 15% or more in humans, for instance - appears to have been incorporated from other organisms, mainly retroviruses (Bestor 2000; Smit 1999). Fortunately, since it does not appear to be active, we can avoid using it.

[40] Indeed, if we find signs of a lack of tree correlation (between trees from different sources) that might indicate such incorporation, this finding would be a subject for further research. (In this particular project, no such finding has been made; however, this question has not been fully examined given the large number of species in the tree and consequent computational time problems plus increased need for lengthy sequence data.)

gene that may have migrated from mitochondria, one may only work (using said gene) with eukaryotes that have that gene exclusively in the nucleus - *if* the migration only happened once in the ancestors of the species in use.

K. Its farthest-apart known structures should be quite far apart (less than 30% identity and/or below 40% with considerable gaps), so that it is not possible to predict the structure of the target by direct homology modeling with any reliability. It should be noted that this criterion is not a necessity for the method to work - it is present for testing the method.

2. Determine sources for phylogenetic sequence data

Need for other proteins

As well as the "central" protein, other sources for sequence data ("other" proteins) are necessary. A phylogeny constructed only using the protein of interest will not be adequately accurate in showing the evolutionary phylogeny of the organisms involved (Bull *et al.* 1993; Cummings, Otto, & Wakeley 1995, 1999; Otto, Cummings, & Wakeley 1996; Russo, Takezaki, & Nei 1996; Zhang, J & Nei 1997)[41]. If we were to build a tree produced solely from the protein of interest, it would likely be incorrect. If so, then we would be predicting past

[41] This problem becomes particularly obvious when one examines the disputes over the human evolutionary tree (Brown, W M *et al.* 1982; Easteal 1990; Gibbs, Collard, & Wood 2000; Glazko & Nei 2003; Kishino & Hasegawa 1989; O'hUigin *et al.* 2002). Another problem with said disputes, admittedly, is probably the short length of time (in evolutionary terms) involved in the divergence in question (which would also minimize the difficulties in any ancestral sequence reconstruction, even if the phylogeny was in error). However, there are a number of problems that emerge with situations of significant divergence, such as long branch attraction (see footnote 52 under "Tree construction methods", on page 27).

ancestral sequences that never existed; such predicted sequences probably would not fold properly[42] or function correctly. If so, then the resulting structure, being incorrect, will make it problematic to then homology model the next structure, even if the sequence for that structure is correct - the incorrect structure will be too far from the correct next structure. We anticipate that only sequences that were along the evolutionary path down and up the tree - that were actually in existence as past sequences for ancestral proteins - will work, since they evidently functioned for organisms in the past. Thus, we will then need to decide on what proteins (and, if necessary, other sequence[43] sources, such as rRNA) are to be used to construct the phylogeny.

---

[42] For instance, such sequences may be trapped into a local (energy) minimum when they tried to fold (or, for our purposes, when they undergo energy minimization as part of homology modeling), being too far from the real sequence.

[43] Of course, rRNA, tRNA, *etc.* also have structures, which should be kept in mind for their usage (Aagaard & Douthwaite 1994; Gutell, Larsen, & Woese 1994; Hancock & Dover 1990; Hickson *et al.* 1996; Kjer 1997; Kraus *et al.* 1992; Morrison & Ellis 1997; Telford, Wise, & Gowri-Shankar 2005; Tillier & Collins 1995; Vawter & Brown 1993; Xia, Xie, & Kjer 2003).

Requirements for other proteins

These sequence sources also need to have several characteristics:

A. The sequences need to be known from as many[44] as possible of the same species as the protein of interest is known in. Preferably, they should (also) be known from species of particular phylogenetic interest. These include (Anderson & Swofford 2004; Gibb *et al.* 2007; Graham, Olmstead, & Barrett 2002; Lartillot, Brinkmann, & Philippe 2007; Moreira, Lopez-Garcia, & Vickerman 2004; Philippe, Lartillot, & Brinkmann 2005) those:
    1. thought to be deeply branching (basal);
    2. that can act as outgroups[45] to species of interest (e.g., those having the central protein);
    3. in some degree of dispute; or
    4. that can act to break up long branches[46].

B. They should likewise be known with a high degree of accuracy. However, this is not as important as for the protein of interest, since errors will be reduced by combining data from multiple sequence sources.[47]

C. The criterion above regarding isozymes can also be relaxed (Arvestad *et al.* 2003). It may be possible to simply remove any cases in which isozymes from the same organism differ significantly from each other, and

[44] Vital enzymes, such as those involved in DNA repair, are particularly likely to be known from multiple organisms, and may moreover be more evolutionarily stable in terms of, for instance, rates of sequence change (Blouin, Butt, & Roger 2005; Knudsen & Miyamoto 2001).
[45] See "Appendix O: Outgroup review/explanation", on page 412.
[46] See footnote 52, on page 27.
[47] This reassurance assumes that we have other sources of information for the phylogenetic relationship of the species from which the erroneous sequence was taken. As long as errors are not systematic - not introducing correlations between sites, for instance - combining data should get around them (Bull *et al.* 1993; Cummings, Otto, & Wakeley 1995, 1999; Otto, Cummings, &

are not clearly distinguishable as to which isozyme corresponds to which isozyme in other organisms (Page 1993); alternatively, we can treat this as polymorphism, the exclusion for which can likewise be relaxed in this usage.[48] (In such a case, randomly choosing which protein to use from each organism would be likely to give a larger evolutionary distance than that actually between the two species. This problem is because one would be counting both the distance between species and the distance between the previously diverged isozymes.)

D. As per ancestral sequence prediction above, we should also be sure to have a sufficient number of different sequence sources that we are able to construct putatively accurate trees using:

    1. a subset of species;
    2. a subset of groups of homologous sequences - which may well be more important, as a form of bootstrapping; and/or

---

Wakeley 1996; Russo, Takezaki, & Nei 1996).

[48] Significant isozyme differences within a species can be defined as those that would make a difference in tree construction depending on which is used. One problem at the point of phylogeny construction comes in if we are using example proteins that have in some species two or more isozymes (e.g., ADH1 Alpha/Beta/Gamma isozymes in primates), or, similarly, significant polymorphism (e.g., for Hemoglobin Alpha). (Another similar case is if using both pseudogenes and active genes was required, although these will have different rates of evolutionary change. The need to use pseudogenes has fortunately not arisen in the present research.) In order to incorporate isozymes, if one can distinguish different "lineages" of said isozymes - e.g., Alpha versus Beta versus Gamma ADH1 in primates (Buhler *et al.* 1984; Cheunq *et al.* 1999) - one can put the protein in question into the tree-building program more than once. For species in which there is only one enzyme (e.g., ADH1 in most mammals), the sequence will be duplicated, whereas in species with more than one isozyme (e.g., primates for ADH1), the individual sequences will each be entered. For polymorphism, or if an isozyme pair is present and active in only one species in the tree (e.g., Hemoglobin Alpha isozymes in *Otolemur crassicaudatus* (Sawada & Schmid 1986)), then (depending, among other considerations, on the program used) it may be necessary to:

- code for this as uncertainty;
- put the species into the tree multiple times (up to once per polymorphic form); or
- do a combination of these (as was done in the present research - see "Usage of polymorphism" on page 64).

3. A combination of the two (a subset of species with a subset of homologous sequences), as has been necessary in the present research.

E. For tree construction, we can use proteins that have not diverged to less than 30% sequence identity, such as eukaryotic cytochrome C (allowing for concerns regarding mitochondrial/nuclear gene migration).

F. We should have structures for purposes of structural alignment, and alignment of sequences to those structures. For the latter alignment to be reliable, the sequences will need to have at least 65% identity to the structural sequences; this level of identity has been shown (Vogt, Etzold, & Argos 1995) to be adequate for sequence alignments to be as valid as structural alignments.

Whether we should be directly using sequence sources other than proteins, such as rRNA and tRNA, is questionable. The primary progress intended in this research concerning tree construction is that of improving our knowledge of protein evolution. It was therefore[49] preferable to use non-protein-coding RNA sequences only indirectly, by using trees derived from it and found in the phylogenetic literature for creating a starting tree (see "3a. Creation of a rough starting tree" on page 72).

[49] For instance, by using protein sequences for the other sequence sources, methods (e.g., alignment) and databases created or adapted for these other sequences can be used for dealing with the central protein.

3a. Creation of a rough starting tree

Tree construction methods

Among methods of tree construction, likelihood methods (of which Bayesian and maximum likelihood methods can be considered a subclass) and, to a lesser degree, distance determination methods are preferable to parsimony, since (among other problems) parsimony (Felsenstein 1978; Hasegawa & Fujiwara 1993; Huelsenbeck 1997; Jin & Nei 1990; Steel & Penny 2000; Yang, Z 1996b; Zhang, J & Nei 1997):

A. is inaccurate on empirical tests;
B. generally lacks compensation for reversion;
C. fails to use all available data[50];
D. is slow;
E. generally lacks a model that is both explicit and biologically reasonable[51]; and
F. tends to produce many apparently-equally-valid trees.

However, likelihood methods can (likewise) be quite slow, particularly for large numbers of species (Felsenstein 1993; Yang, Z 1994, 2000a). In essence,

[50] Parsimony methods - other than weighted parsimony (Felsenstein 1981) - fail to take into account information as to the likelihood of changes. For instance, the substitution of a tryptophan for an alanine in a protein's active site is rather less likely to be evolutionarily accepted than the substitution of an aspartic acid for a glutamic acid in a surface loop.

[51] Not all sequence locations are under evolutionary constraints such that changes are unlikely to be accepted. Moreover, polymorphism even on locations that are the subject of selection does happen within species. Both of these indicate that the only explicit model generally used in parsimony - that all mutations are unlikely - is unreasonable. Indeed, over a sufficiently long evolutionary timespan, it would be anticipated that all locations not extremely constrained, by functional constraints always in place for the protein or RNA in question, would eventually change. Note that if a substitution matrix is multiplied by itself sufficient times - as was done to construct the PAM series of matrices (Dayhoff, Schwartz, & Orcutt 1978) - the likelihood that an amino acid will remain unchanged will eventually go below 50%.

maximum likelihood involves iterated estimation of the most likely branch lengths - and often other parameters (Yang, Z 2000b) - for each possible tree, followed by the selection of the most likely tree among those examined. One technique ("quartet puzzling") that has been implemented (von Haeseler & Strimmer 2003; Schmidt *et al.* 2002; Strimmer & von Haeseler 1999) to compensate for this time requirement is looking at only four species at a time, then putting together the resulting trees into an entire tree. In some respects, this is a compromise between distance methods (which essentially examine only two species at once) and full maximum likelihood. The latter derives much of its accuracy from examining distances from ancestral nodes. This examination is not done at all for distance methods[52] and is only done to a limited degree for the "quartet puzzling" method (von Haeseler & Strimmer 2003; Schmidt *et al.* 2002; Strimmer & von Haeseler 1996; Strimmer, Goldman, & von Haeseler 1997; Strimmer & von Haeseler 1999). However, one problem with distance methods (and with "quartet

[52] In distance methods, all distances examined are between end sequences; this is part of what makes the "long branch attraction" problem worse for distance methods (Huelsenbeck 1997), although it is certainly seen in other methods (Anderson & Swofford 2004; Felsenstein 1978; Lartillot, Brinkmann, & Philippe 2007; Ranwez & Gascuel 2001), as is seen in the present study. In this phenomenon, species that are highly divergent are attracted to each other in the tree (i.e., branch more closely - either all in one apparent group (clade), or further down in the tree) to one another than occurs in the true phylogeny); note that this is a difference in terms of branching order, not simply the distances themselves. Most commonly, long-branching species will be attracted to other long-branching species/groups that are close to the root of the tree - e.g., a long-branching eukaryota may be attracted to archaea, as seen in “Tree search with Eukaryota (subset)”, on page 300; also see “Appendix O: Outgroup review/explanation”, on page 412. (One reason for this can be that, by chance, species that have had many neutral mutations will have a closer amino acid (or nucleotide) distribution in various portions of their sequences than will species under more selective pressure for the locations in question.) It is sometimes helpful to add species that break up long branches (Anderson & Swofford 2004; Gibb *et al.* 2007; Lartillot, Brinkmann, & Philippe 2007; Moreira, Lopez-Garcia, & Vickerman 2004; Philippe, Lartillot, & Brinkmann 2005). Correlations of mutations (as, for instance, MrBayes attempts to handle (to a mild degree) via the “covarion” option - see footnote 200 under “MrBayes code alterations”, on page 99) can also cause problems with the placement of species with long branches (Gaucher, Miyamoto, & Benner 2001; Gaucher *et al.* 2002; Lopez, Forterre, & Philippe 1999). Similarly, the differences in selective pressure between different locations can cause problems (Lartillot,

puzzling") is that they have difficulties with species that lack sequences in common[53]. Maximum likelihood also appears to be more robust than distance determination (Cunningham, Zhu, & Hillis 1998; Huelsenbeck 1995; Tillier & Collins 1995; Yang, Z 1994). relative to potential problems such as correlations between sites. Moreover, maximum likelihood methods allow for the determination of the optimal parameters in other respects, such as estimated rate variation.

Need for starting tree

Given the slowness of maximum likelihood (and of Bayesian techniques) for trees involving considerable numbers of species, as is likely to be the case for a tree involving vital proteins with significant divergences in sequences, it is necessary to find ways to speed up tree determination. With regard to Bayesian techniques, one way to increase their speed is to put in an initial starting tree that is, while not assumed to be entirely correct, closer to the correct tree than the randomly-generated tree (to be later rearranged) that they would otherwise start

---

Brinkmann, & Philippe 2007).

[53] For instance, one can use myoglobin for a protein sequence source only for species with muscles, causing a potential problem with distance methods and "quartet puzzling". (This problem was encountered in the present research. For instance, the results from Tree-Puzzle (von Haeseler & Strimmer 2003; Schmidt *et al.* 2002; Strimmer & von Haeseler 1999) for Bacteria showed distances negatively correlated with the number of genes the organisms had in common, with a resulting tree that was more a source of amusement for at least one microbiologist than anything useful. However, quartet puzzling was found useful for one subset of the species (Archaea) that had significant sequence data in common.) Distance methods use a distance matrix - a table of distances from one species to another - for the basis for constructing a tree; there would be a lack of data for any pair of species without genes in common. Quartet puzzling has problems with such missing data due to it both:

- Requiring each of the 4 species used in a quartet to have sequences in common; and
- Using distance determination for the determination of properties such as rate variation (Strimmer 1997) - this is problematic due to the need for a full distance matrix.

An examination of the code in Tree-Puzzle was done in an attempt to cure the first of these problems, but was determined not to be practical in consideration of the time limits on this

with (Huelsenbeck & Ronquist 2001; Huelsenbeck *et al.* 2006). Another method of speeding up tree determination is to split up the tree into smaller sections; however, the best available method for determining how to do this, Rec-I-DCM3, likewise requires a starting tree (Huson, Nettles, & Warnow 1999; Roshan *et al.* 2004a; Roshan *et al.* 2004b). In constructing such a starting tree, it is necessary to avoid usage of phylogenies produced by the phenetic method (Sneath & Sokal 1973), which groups by similarity of (primarily morphological) characteristics[54], not ancestry. Instead, one should use those produced by the cladistic method (Hennig 1979), which groups by ancestry, not similarity of characteristics (Futuyma 1986). This usage of previously constructed trees does have the advantage of incorporating information not otherwise contained in the sequences examined in a particular study. These sources range from rRNA (for a study such as this one concentrating on proteins) to fossil evidence (see footnote 54).

3b. Alignment of other sequences

Multiple alignments

One known problem in phylogenetic estimation is the dependence of the results on the sequence alignment used (Kjer 1995; Lake 1991; Morrison & Ellis 1997).

---

research, especially considering that the second problem would still be present.

[54] This stipulation does not mean that morphological characteristics should be discarded - among other reasons, they are the best means of inserting the most available fossil information (Wagner 2000) into a tree - but grouping solely by them is not, in most cases, grouping by ancestry. (Morphological characteristics involved in reproduction may be exceptions; if two fossilized mammals or birds appear incapable of mating due to, e.g., size differences, then they can be deduced to be different species, for instance.) Morphological (and behavioral, for which mating determinants are particularly of interest) characteristics appear to work best as confirmation of trees derived from other data; where they are incongruent, there is reason to investigate further (Swofford 1991; Wyss, Novacek, & McKenna 1987; Xia, Xie, & Kjer 2003).

If a progressive multiple sequence alignment[55] is created via a guide tree - the method (with a distance-based guide tree) used by such programs as ClustalW (Thompson, J D, Higgins, & Gibson 1994) - this results in the guide tree biasing any trees derived from the aligned sequences (Lake 1991). The best - or at least most applicable for our purposes - method known to solve this problem is to use structurally based alignments as much as possible (Kjer 1995; Morrison & Ellis 1997; Xia, Xie, & Kjer 2003). The basis of this method is that proteins (like ribosomal RNA, for which the method was originally developed) conserve structure more than sequence, particularly when function is conserved (Rossmann, Moras, & Olsen 1974). Sufficiently close conservation[56] allows the usage of a star tree in alignments (aligning other (sufficiently close) sequences of a given protein to one central sequence - i.e., to one with a known structure), which minimizes the bias due to the guide tree (Lake 1991).

Structural alignments

A variety of means are available to perform protein structural alignments (Falicov & Cohen 1996; Gerstein & Levitt 1996; Levitt & Gerstein 1998; Yang, A-S & Honig 2000a, 2000b, 2000c; Zemla *et al.* 1997). The most widely accepted measure of structural deviation (including as the basis, or at least a basis, for several other measures) is the Root Mean Square Deviation (RMSD; the square root of the mean of the squared distances between the aligned atoms); methods

[55] A multiple sequence alignment is necessary for Bayesian (including maximum likelihood) phylogenetics (and for parsimony), and is preferable for internal consistency with distance-based methods.

[56] I.e., close enough that the usage of a variety of alignment sources (including both matrices and structural alignment) should yield essentially the same alignment. For protein sequences, this is

more directly attempting to minimize this (e.g., Structal (Gerstein & Levitt 1996, 1998)) therefore appear preferable.

Sequence alignments

We must decide what matrices, gap penalties, *etc.* are to be used for the alignment of sequences to structures. Previously, much work has been done in this area; however, most of it is on the topic of searching databases using local alignment (Altschul *et al.* 1990; Henikoff & Henikoff 2000). Local alignment is not suitable for our purposes, since it concentrates on a similar subsection of the sequences to be aligned (Altschul *et al.* 1990; Saqi, Russell, & Sternberg 1998; Vogt, Etzold, & Argos 1995). We are not looking only at the similar regions of the proteins of interest, but at the entire protein.

However, there is information available on optimal alignment conditions (matrices and gap penalties) for the congruence of alignments with manual alignments of well-understood proteins and structural alignments of proteins with known 3D-structure (Abagyan & Batalov 1997; Burke *et al.* 1999; Domingues *et al.* 2000; Johnson, M S & Overington 1993; Vogt, Etzold, & Argos 1995). Since a brute-force search of the entire possible range of matrices and/or gap penalties is impractical due to computation time, using previous research such as this to determine the values to start out with will be necessary. The matrices, or at least one of the matrices, selected at this step may be used later for tree

---

generally true at or above 65% identity (Vogt, Etzold, & Argos 1995).

determination[57], or a matrix (e.g., WAG (Whelan & Goldman 2001)) created specifically for phylogenetic usage[58] may be used.

4. Tree refinement

This stage consists of two aspects:

A. The correction of the structure of the starting tree, which may be arbitrary in some locations or otherwise likely to be inaccurate;
B. The original estimation (or refinement, at later points) of distances, which are needed for REC-I-DCM3[59] (Roshan *et al.* 2004a; Roshan *et al.* 2004b) and for ancestral sequence determination (to determine the likelihood of mutations happening between two nodes).

These will use data from the other sequences and, once they are aligned, the central sequences.

5. Alignment of central sequences

The next stage, for which the tree will hopefully be helpful (e.g., to determine relative species weights (Altschul, Carroll, & Lipman 1989; Felsenstein 1973, 1985b)), is to align the central (main) sequences. For those sequences close to a usable (non-target) structure, this step can be the same (albeit with further manual checking) as with the other sequences. However, once the sequence identity between a usable structure and a sequence of interest has dropped

[57] That is, by methods other than parsimony.
[58] This criterion/description includes varying it for different amino acid compositions, which is frequently not done for matrices used for alignment, as opposed to searches (Schaffer *et al.* 2001). See also footnote 221 under “Partitions: State frequencies”, on page 107.

below 65% (Vogt, Etzold, & Argos 1995), means that are more accurate become necessary. If multiple matrices (ideally, matrices derived from multiple sources/methodologies) give the same result for an alignment of a sequence, then it may be trustworthy to use that alignment despite a below-65% identity. However, it is unfortunately unlikely that this agreement will happen with low percent identity sequences. Among the possible solutions at this stage are:

- threading[60] (Yang, A-S 2002; Zhang, K Y J & Eisenberg 1994);
- the usage - similarly to Pfam (Bateman *et al.* 2002) - of a Hidden Markov Model (HMM) created using the existing alignment (Eddy 1995); or
- some combination of these - e.g., the usage of a HMM with information from both sequences and (known and possibly predicted) structural aspects, such as secondary structure and accessibility (Elofsson 2002; Goldman, Thorne, & Jones 1996; Koshi & Goldstein 1995).

With the alignment of the central sequences, some degree of tree refinement is possible and, in two cases/aspects, preferable:

1. In order to determine distances better, so that they will more closely correspond to the degree of evolutionary changes in the central sequences;
2. For any organisms for which one does not have other sequences, or has an inadequate number of other sequences, although this is only advisable

[59] See under "Need for starting tree", on page 31.
[60] However, threading is sometimes found to be better at recognizing folds than it is at doing sequence/structure alignments (Lathrop *et al.* 1998; Mirny & Shakhnovich 1998). See also footnote 9 under "2. Phylogenetics - Ancestral Sequence Prediction", on page 7.

if one has other data[61] indicating the proper placement (albeit initially without distances) of an organism on the tree.

6. Determination of ancestral sequences

Some types of tree construction - namely parsimony and Bayesian methods (including maximum likelihood) - generate an ancestral sequence (or a set of possible ancestral sequences) for each ancestral node as a part of their construction (Hartigan 1973; Higgins 2000; Yang, Z, Kumar, & Nei 1995). The distance method of tree construction does not, but trees constructed using it can be used, along with estimates of a particular mutation's likelihood, to create such predictions. Predictions of ancestral sequences can be created after the construction of a tree by any of the methods, even if the method inherently creates ancestral sequence predictions while constructing the tree. This property is of advantage if one wishes to take advantage of more information in generating the prediction than is practically usable in generating the tree - e.g., more species or high-complexity structural information - or, as in the present work, if one needs to narrow down the predicted possible sequences using such information[62].

[61] Among the sources of other data are gene splits (Stechmann & Cavalier-Smith 2003), fusions (Stechmann & Cavalier-Smith 2002), and gaps/insertions (Baldauf & Palmer 1993).

[62] It is possible to put predicted sequences in a tree (including into the same tree used to predict them, although this will only be of value for maximum likelihood or parsimony if other information is used to narrow down the possible sequences) in order to use them for assistance in tree topology determination and/or tree distance determination. This has been done in the present work.

If using a parsimony or likelihood (including Bayesian and maximum likelihood) method of tree determination, the extraction of probable ancestral sequences at this stage is relatively straightforward, with three caveats:

1. One has to decide whether one creates an ancestral sequence for each node (intersection of tree branches), which may increase accuracy, or skips some nodes, which may increase speed. If the ancestral sequence is at all likely to change significantly relative to the sequence of the prior structures (including structural models), such as with the introduction of gaps or insertions or significant changes in amino acids[63], then it appears advisable to try an ancestral sequence reconstruction for a node. If it turns out that the node's sequence does not differ significantly, one can skip model building and go to the next node for the construction of a further-away ancestral sequence.
2. One is likely to get, not absolute sequences, but probabilities for the possible amino acids present in a given ancestral sequence. Prior structural information may sometimes be able to indicate what sequence(s) among the possibilities should be tried (Aszodi, Munro, & Taylor 1997; Azarya-Sprinzak *et al.* 1997; Blundell 1991; Bowie, Luthy, & Eisenberg 1991; Fornasari, Parisi, & Echave 2002; Koehl & Levitt 2002; Ponder & Richards 1987a, 1987b; Sunyaev *et al.* 1997; Wilmanns & Eisenberg 1995; Word *et al.* 2000). Other sources for this information include the

[63] E.g., the introduction or removal of a proline or glycine, or a significant change in volume or hydrophobicity/charge.

examination of correlations between residues[64], which may indicate that some combinations of amino acids are unlikely (Fukami-Kobayashi, Schreiber, & Benner 2002; Gobel *et al.* 1994; Pollock, Taylor, & Goldman 1999; Saraf, Moore, & Maranas 2003; Singer, Vriend, & Bywater 2002).

3. The determination of gaps as being present in particular locations - i.e., what residues are likely to be insertions in the present-day sequences relative to the ancestral sequences, and what residues are likely to be deletions (from the ancestral sequences) in a subset of the present-day sequences[65]. This problem is largely unsolved, although some heuristics[66] exist. This difficulty is at least partially[67] because much more is known about the likelihood of residues to mutate to other residues (as expressed in matrices) than about the likelihood of insertions and deletions (as can be seen in the arbitrary nature of many gap models, as noted under "2. Phylogenetics - Ancestral Sequence Prediction", on page 8). An examination of the DNA sequences may be of use in some evolutionarily recent cases (Chang, M S S & Benner 2004).

---

[64] This would particularly include residues close by each other in 3D space and/or that influence each other via movements of secondary structures, the latter of which is rather more difficult to analyze.

[65] If the residues in question were deleted from all present-day sequences, then it would be very difficult, if not impossible, to reconstruct them from the present-day sequences.

[66] These heuristics are primarily via representing gaps as binary characters then using parsimony (Edwards, R J & Shields 2004) or approximate Bayesian/maximum likelihood methods (Huelsenbeck *et al.* 2006).

[67] Another difficulty may be that thinking (and writing, as can perhaps be seen in the present work) about gaps/insertions/deletions ("indels") is more difficult than thinking (and writing) about amino acids in protein sequences, perhaps partially because indels only make sense in the context of an alignment.

Ultimately, one may need to build multiple models (from different sequences) and use them to help eliminate some sequences as unlikely (see "8. Examination of models" on page 41).

7. Model building

Given an alignment between known (3D) structures and a sequence to be modeled, the primary stages of homology modeling are as follows:

A. Assignment of initial coordinates from existing structures (templates; these can be experimentally determined structures, prior models, or both). This should include backbone coordinates in any location without an insertion, deletion, or (possibly) mutation to/from a proline or glycine. It should likewise include sidechain coordinates if the residue is not mutated, and in some cases even if it is. (E.g., one can derive the coordinates for phenylalanine from those for tyrosine, by removing the sidechain OH; even if this is not completely physically realistic due to the possible hydrogen bonds from the OH, it should be at least a good starting point, and help one avoid the otherwise necessary loop or rotamer search.)

B. Loop searches (of known 3D structures) for areas[68] not found in the template. In these, sections of other structures are aligned to (i.e., superimposed on) the surrounding existing template residues and the otherwise-missing residues/sidechains are replaced (or averaged, for the existing template residues).

[68] This is done for at least the backbone coordinates, and ideally the sidechains (Chakrabarti & Pal 2001; Shortle 2002), to eliminate the need for a rotamer library search (see item C).

C. Usage of a rotamer library/search[69] to derive any remaining (not derivable from the templates or loop searches) residue sidechains.

D. If loop searches, deletions, or other significant changes (e.g., from/to proline or glycine) have been necessary, "vacuum"/"dry"[70] energy minimization[71] with residues that neither were altered nor are near[72] altered residues being "frozen" (kept in the same place) or, at least, restrained in movement.

E. Vacuum energy minimization without freezing any parts of the molecule.

F. The addition of the solvent (water, for the present research).

G. The energy minimization of water, while keeping most or all of the protein frozen.

H. "Wet"/"Full" (with water surrounding it) energy minimization of the protein; water very far from it (such as in the corners of the simulation "box") may be frozen.

I. If areas of the model appear to be stuck in a local minimum[73], simulated annealing. In this process, molecular movement with an increased simulated “temperature” is used to "shake and bake" the structure - or

---

[69] A rotamer search would be usually within the angle ranges found in a rotamer library (Lovell *et al.* 1999, 2000; Word *et al.* 1999b).

[70] By “vacuum”/”dry” is meant without water, and possibly with no explicit charge interactions (i.e., Coulomb's law interactions ignored).

[71] I.e., doing alterations to reduce the predicted potential energy. These generally start with alterations of moderate size, then increase the size of these if the alterations so far are succeeding or decrease the size of these (and try again) if the alterations are not succeeding (i.e., appear to have overshot). The exact algorithm for this is controlled by the selection of a “minimizer”.

[72] By “near” is meant “near in sequence”. The question of whether to “unfreeze” residues nearby to altered ones in the templates’ 3D structure is an excellent one. This procedure would ideally be done, but may lead to too few atoms being “frozen” for said “freezing” to accomplish anything.

[73] Among the means of detecting a local minimum are if the residues are in a biologically unlikely state from enzymatic considerations, rotameric or Ramachandran (backbone) configurations, or

possibly only the problematic portions (Flohil, Vriend, & Berendsen 2002) - so that it goes into another state; this is followed by (further) energy minimization.

If, in the above process, one is modeling more than one sequence (including both differing amino acids and differing gaps), and some of the models encounter significant problems[74], then this may indicate that the corresponding sequences are not realistic (see "8. Examination of models", below, for how this information may be used).

8. Examination of models

The resulting model(s) can then be examined for likely levels of realism and of errors. For multiple models of the same sequence (as may be derived from different model building conditions), one will wish to determine which portions are the most reliable for each model, to preferentially use these in building the next model. If one has multiple sequences, then the models may assist one (such as by noting overcrowding, which would tend to indicate that, for instance, a tyrosine was more likely than a tryptophan, or a valine than an isoleucine) in determining which sequence(s) are more realistic. Whatever sequence(s) are determined to be realistic may then be put into the tree very close to the ancestral node, to simulate an ancestral species branching off at that point, and the sequence(s) may be put into the alignment (as has been done in the present research), to

---

examination of similar proteins.

[74] Examples of such include severely overlapping Van der Waals radii not relievable by energy minimization or a backbone configuration needed that cannot be found by a loop search.

assist in the alignment[75] of more sequences. Those of particular importance include those nearer the target structure[76], and thus without structures usable to align them.

Ultimately, when the sequence reached is not an ancestral sequence, but that of a target organism, this stage would be[77] the conclusion of the research. After going through the possible chains of homology models, we will then analyze what produced correct results and what did not. (Correct results, for the best realistically achievable case, can be defined[78] as those that are superimposable on (structurally alignable to) the actual target structure to within the resolution[79] of that structure.)

---

[75] We have not aligned all of the central protein's sequences at once; we have instead added some sequences only after modeling and adding ancestral sequences, to improve the alignment.

[76] This includes any sequences that are nearer the target structure than to the templates used for producing the initial models. In other words, these sequences are the descendents of the ancestral sequences to be modeled as we go back up the tree toward the target structure.

[77] In the case of the present research, this stage has not been reached, although considerable progress has been made toward it.

[78] If there is more than one structure known for the target sequence (as is the case in this research), then an additional comparison can be with how well these structures can be superimposed on each other. See footnote 22 under "3. Homology Modeling", on page 13, for another definition. Note that even the latter (more relaxed) definition is not matchable, as far as we are aware, by any *de novo*/*ab initio* method for a protein of the size of the central protein chosen for the present research.

[79] This criterion is as per the evaluation of modeling in CASP1-3 (Gerstein & Levitt 1998; Yang, A-S & Honig 2000a; Zemla *et al.* 1999).

Chapter 3: Detailed Materials and Methods

Choice and Availability of Programs and Data

Open-source (Coar 2006)[80] programs have been used[81] whenever possible, in order to:

1. Increase the reproducibility of these results, including with regard to any needed modifications of programs; and
2. Make sure that all algorithmic methods used are properly published[82].

When this has not been possible, then programs have been selected for being freely available, with source code, and likely to be available to be placed under an open-source license if necessary (e.g., those available via http://kinemage.biochem.due.edu/). When the name of a program is such that it may be confusing, it has been printed in a different font (e.g., "`reduce`").

[80] Note, incidentally, that open-source does not mean public domain; nor does it mean "freeware". Open-source programs come under a license such that it is not possible to redistribute either that program (including via using it online, for the AGPL used for programs created in this work), or a modified version of it, without likewise distributing the source code.

[81] Programs that have been used for such ancillary functions as molecular display or statistical analysis may be exceptions to this, but are not of sufficient importance to be cited (any more than the word processor on which this document is written will be cited), not being of significance for the reproduction of this work. An exception regarding ancillary functions would be a program like KiNG (Richardson, D C 2007), used for some molecular display, that is also associated with other functions, e.g., MolProbity, used to evaluate models (Davis *et al.* 2007; Lovell *et al.* 2003).

[82] A prior graduate student in our laboratory (Drennan, Richards, & Kahn 1993; Drennan 2001) previously encountered problems of this nature with a program, DSSP (Kabsch & Sander 1983), which is commonly used to classify the secondary structures of known protein structures. The program determines the beginnings and ends of alpha helices as being one residue shorter at each end than the published algorithm does; the authors of DSSP acknowledge the difference privately (or at least did as of 1988), but have not published this information. The program is not freely available with source code redistributable; it is thus not possible for others - e.g., our laboratory (Drennan, Richards, & Kahn 1993) - to rectify this situation for others. Partially in response, all consideration of secondary structure in experimental structures for this project has used the records in the source PDB files, which are not *necessarily* derived from DSSP. Some older structures, in particular, appear to have had their secondary structures determined by other methods, resulting in configurations not capable of output by DSSP; these include overlaps between alpha-helix and beta-sheet, for instance, in the main chicken DHFR structure used,

Programs not created locally that were used in this research or mentioned in this dissertation are listed in “Appendix Q: Non-local programs used/mentioned”, on page 423. All other programs used for citable purposes (i.e., not for uses similar to word processing) were created locally.

Similarly, all external databases used are publicly (and freely) available (and all generated databases are publicly available without fee, provided the licensing restrictions are obeyed; see below). Most importantly, experimentally determined (protein) structures not in the PDB (Protein Data Bank; (Berman *et al.* 2000)) were ignored, as were papers depending on or otherwise heavily citing such sources, including indirectly. Sequences not deposited in GenBank (Benson *et al.* 2000) have likewise not been used, nor have papers using such sequences.[83].

When a choice of different papers giving information on a topic (particularly applicable for reviews) was available, freely available online sources (Suber 2007) have been used when possible. Moreover, it is intended that journal publications emerging from this dissertation will be in journals that are freely available (either immediately or after some reasonable period).

8DFR (McTique *et al.* 1993).

[83] Note that most journals today will not accept for publication material depending on sequences not in GenBank (or protein structures not in the PDB). Material such as program source code (for a methodology publication in particular) and databases appear ethically equivalent to sequence or structural data, in that they can be provided without reduction in the material available for the author’s further research (in contrast to, for instance, tissue samples).

Programs created or modified in this work are available as supplemental datafiles to this dissertation and/or can be downloaded, as specified below (generally from http://cesario.rutgers.edu/easmith/research/ or pages linked from there - see “Appendix R: Supplemental files and URLs” on page 426). All programs created or significantly[84] modified in this work are available under an open source license; for programs created for this work, the license is the AGPL (Foundation 2007), version 3. The 180 programs created *de novo* for this work (see “Appendix P: Perl programs created”, on page 415) are written in a programming language, Perl (Wall, Christiansen, & Orwant 2000), that is freely available. These are indicated with a file ending of ".pl"; any program with a filename ending in ".pl" was created locally, and is available under the AGPL, version 3. All Perl modules[85] - e.g., for phylogenetic work (Vos 2006) - used are also freely available from CPAN, the Comprehensive Perl Archive Network (see http://www.cpan.org). Datasets/databases created in this work can likewise be downloaded as specified below. The databases in question, like this dissertation, are licensed[86] under a Creative Commons Attribution-ShareAlike 2.5 License (Commons, C 2006).

---

[84] By "significantly", we mean beyond those changes needed to achieve compilation on and display formatting suitable for the local systems, plus, for PHYLIP (Felsenstein 1993), as directed in the documentation as necessary to alter various built-in limits.

[85] These are external libraries of subroutines for Perl programs.

[86] However, factual material contained in said databases is (fortunately) not subject to copyright law in the United States, only the compilation of said factual material (Commons, S 2007). The proper attribution/acknowledgement format/means to satisfy the Creative Commons license’s requirements in that regard is the normal academic/scientific citation (of papers emerging from this research and/or of the dissertation itself; the first is preferred, if possible, being easier to track citations of). Note also that the Creative Commons license in use is not one such as is customary in Public Library of Science articles (see http://www.plos.org/oa/definition.html), but contains limitations on the distribution of derivative works. Such works may only be distributed, including via scientific publications unless they fall under “fair use” rights, if they are placed under the same license.

Modifications to some programs and other files are available in the form of "patchfiles", applicable to the programs/datafiles by the UNIX command "`patch`" using the "-El" options. These, listed in "Appendix Q: Non-local programs used/mentioned", on page 423, are available[87] under http://cesario.rutgers.edu/easmith/research/patches/ with a ".patch" ending, and will be brought to the attention of the original authors of the programs, as applicable.

Another selection criterion for the programs and methods used is that those programs/methods created using data from the target protein structures (e.g., modeling programs based on databases that include the target protein structures) were avoided when possible. (The sole exception will be those used for evaluation of the final models via comparison to the physically known target protein structures; see "Final evaluation", on page 356.) Such programs/methods were completely avoided if either:

A. the target protein structures might have made a significant influence on the database, due to their proportion to other protein structures (e.g., the Pfam alignment for the target protein was not used, since it may have been partially constructed using the target protein structures); or

[87] The terms under which Rutgers requires the release of this dissertation do not appear to allow for the release of patches (as supplemental files) to material copyrighted by others, even under GNU or similar open-source copyrights.

B. the effects of the usage of the target protein structures on the program/method were obscure[88] or could not otherwise be ruled out as affecting the results. (For instance, loop searches were done using locally created programs, not existing ones that could pick up loop structures from target structures).

Similarly, papers noted as using the target protein structures will not be read until the final evaluation stage (see “Final evaluation”, on page 356).

Methods and Data

1. Determination of central protein

Potential candidate proteins were initially located by examining metabolic pathways for vital enzymes found in most organisms, particularly eukaryota (given, among other considerations the difficulty in determining well-defined species in non-eukaryotes). Proteins were then selected from among these based on the other criteria, in particular:

- not being either:
  - mitochondrial or
  - likely to have migrated from mitochondria since the origin of eukaryotes (Bensasson, Zhang, & Hewitt 2000; Berg & Kurland 2000; Shafer *et al.* 1999);

[88] Instances of potentially “obscure” program methodologies include:
- Those in programs that are not completely open-source, such as Modeller (Sali & Blundell 1993; Sali & Overington 1994; Sali 1995, 2001); and
- Those using neural networks (Fariselli *et al.* 2001; Lundstrom *et al.* 2001; Wallner & Elofsson 2003) and similar "black box" techniques.

- not being membrane proteins;
- not being found in multiple different copies[89]
- not being multisubunit proteins[90]; and
- being found in the PDB with structures from multiple, widely-separated species at a low percentage identity

Central protein candidates

Two primary candidates were found:

1. Orotidine 5'-phosphate decarboxylase (Bell & Jones 1991; Cui *et al.* 1999; Ghim, Nielsen, & Neuhard 1994; Kimsey & Kaiser 1992; Miller *et al.* 1999; Ohmstede *et al.* 1986; Radford 1993; Sakai, Kazarimoto, & Tani 1991; Strych, Wohlfarth, & Winkler 1994; Suchi 1988; Turnbough *et al.* 1987; Yaoi *et al.* 2000), abbreviated ORO (EC 4.1.1.23). This enzyme catalyzes the last stage of pyrimidine nucleotide biosynthesis, from orotidine 5'-monophosphate to uridine 5'-monophosphate (UMP). Structures of it are known from eukaryota (*S. cerevisiae*), bacteria (*E. coli* and *Bacillus subtilis*), and archaea (*Methanobacterium thermoautotrophicum*). It exists in all species except for closely parasitic ones (viruses and other obligate intracellular pathogens). Partially due to its extreme efficiency of function[91], it has been extensively studied (Acheson *et al.* 1990; Appleby *et al.* 2000; Begley, Appleby, & Ealick 2000; Krungkrai *et al.* 2001; Lee, J K & Houk

[89] Copies due to ploidy were not considered problematic. In other words, we did not consider the "duplication" of genes on homologous chromosomes (or, if no real divergence had taken place, duplicate chromosomes from whole-genome duplications) to be problematic.
[90] For instance, many enzymes involved in DNA replication, while otherwise highly suitable due to

1997; Miller *et al.* 2000a; Miller *et al.* 2000b; Miller *et al.* 2000c; Porter & Short 2000; Rishavy & Cleland 2000; Smiley *et al.* 1991; Warshel *et al.* 2000; Wu, N *et al.* 2000a; Wu, N *et al.* 2000b). One difficulty with ORO as a target protein is that in known metazoa and some other species (e.g., *D. discoideum*), it is fused (Yablonski *et al.* 1996) with the enzyme catalyzing the prior stage in pyrimidine nucleotide biosynthesis (orotate phosphoribosyl-transferase); it also has a large number of insertions[92] in various species (Traut & Temple 2000). Moreover, it is normally functional as a dimer, with possibly asymmetric interactions between the two monomers (Harris *et al.* 2000).

2. Dihydrofolate reductase (or tetrahydrofolate dehydrogenase), abbreviated DHFR (EC 1.5.1.3), which catalyzes the reduction (addition of hydrogens) to dihydrofolate to form tetrahydrofolate (and in some cases the reduction of folate to dihydrofolate), an essential reaction in the thymidylate (dTMP) synthesis pathway. DHFR has been extensively studied (a search on PubMed for '"dihydrofolate reductase" OR "tetrahydrofolate dehydrogenase"' currently (as of Jan 7, 2008) indicates 5925 articles); this discussion will accordingly focus on only those aspects of relevance to the present research. Given its necessity for correct DNA replication[93], it is perhaps unsurprising that DHFR has been frequently targeted for inhibition

---

their vital nature, are unfortunately multisubunit.

[91] ORO has the highest degree of rate enhancement of a reaction known for any enzyme.

[92] While the presence or absence of these insertions is potentially valuable for phylogenetic purposes (Baldauf & Palmer 1993; Stechmann & Cavalier-Smith 2002), they would be troublesome for modeling purposes (necessitating multiple loop searches over and above those desirable or needed for other reasons).

[93] A lack of dTMP causes either cell cycle arrest or disruption of DNA synthesis.

in bacterial infections (e.g., by trimethoprim (Krahn *et al.* 2007)), eukaryotic infections (e.g., versus malaria (Gregson & Plowe 2005)), and cancer (e.g., by methotrexate (Goodsell 1999)). Structures of DHFR are known from metazoa (including humans and mice), fungi, several other eukaryotic species (two types of human malaria, *Plasmodium falciparum* and *Plasmodium vivax*, and two parasites causing diarrhea, *Cryptosporidium hominis* and *Cryptosporidium parvum*), and multiple bacteria[94]); please see Figure 3.1, on page 52. One difficulty[95] is that DHFR in eukaryotes other than fungi and metazoa (Stechmann & Cavalier-Smith 2002, 2003) is fused with another enzyme in the thymidylate synthesis pathway, the eponymous thymidylate synthase (TS; EC 2.1.1.45), which catalyzes the synthesis of dTMP from dUMP using hydrogens from tetrahydrofolate; the fused enzyme also appears to be dimeric (Shallom *et al.* 1999). However, the metazoan and fungal DHFRs are sufficiently different from each other (below 40% identity[96], with significant gaps) to be usable as source/target proteins, thus hopefully allowing avoiding venturing into modeling the fused structures.

Ultimately, DHFR was chosen, due both to the concerns noted above with regard to ORO's fusion and dimerization and to the difficulty of phylogenetic tree determination necessarily involving all three "superkingdoms" (eukaryota,

[94] Note that some bacterial DHFRs appear to be unrelated and are plasmid-borne (Krahn *et al.* 2007).
[95] At least, (potential) difficulty for modeling as opposed to phylogenetic purposes (Stechmann & Cavalier-Smith 2002, 2003).
[96] The identity is 31% by a BLOSUM62 (Henikoff & Henikoff 1992) alignment with the default settings for ClustalW at EBI (http://www.ebi.ac.uk).

archaea, and bacteria)[97], especially considering the probable symbiotic origin of eukaryotes (Brown, J R & Doolittle 1997; Margulis 1996). However, ORO was kept as an additional source of sequence data for tree determination (see "Other proteins used" on page 58).

Selection of structures and other sequences

Structures for DHFR were required to have backbones and side chains (not be alpha carbon only) with a resolution of 3 Ang. or less; NMR structures were avoided when possible. After locating the initial DHFR, DHFR/TS, and TS structures, the sequences from the residues seen in the PDB file ATOM (coordinate) records were determined, using both ASTRAL (Brenner, Koehl, & Levitt 2000; Brenner *et al.* 2000) and a locally-written program, "extract.atom.seqs.pl" to determine this. Residues in the PDB file SEQRES sequence, but not actually seen in the ATOM records, were either replaced with "x"es or, if at the ends, trimmed off (as were any initiating methionines, His tags, *etc.*)[98]. Searches were then done using NCBI's `blastp` (Altschul *et al.* 1990; Altschul *et al.* 1997; Gertz 2006) to locate further structures (with manual verification); the default settings (e.g., using the BLOSUM62 matrix) were used, aside from increasing the default gap existence penalty[99]. The minimum percent

[97] The consideration of all three would be needed so that there would be an “outgroup” (e.g., bacteria) for the ancestor of the other two groups (e.g., archaea and eukaryota). Please see “Appendix O: Outgroup review/explanation”, on page 412.

[98] No editing to match the sequence found in the PDB file ATOM records was done on the sequences (DHFR or otherwise) used for phylogenetic work, however, except that initiating methionines, his tags, and other non-native alterations were removed (which in some cases (e.g., 1U70 chain A) meant using a different (non-mutant) sequence than that in the ATOM records).

[99] The default penalty is 11. The penalty was raised by 1, to (the maximum) of 12, to discourage unnecessary gaps, since gaps are likely to be problematic (Golubchik *et al.* 2007).

identity considered usable was 30% (Rost 1999). Following this search, a similar search[100] was done versus the NCBI nr sequence database (Wheeler *et al.* 2000). Sequences with 65% or more identity were preferred (see item F under "Requirements for other proteins", on page 27) but not required (although above a 30% identity *was* required), given the greater degree of manual intervention and other methods expected to extend the usable range of DHFR/TS alignments.

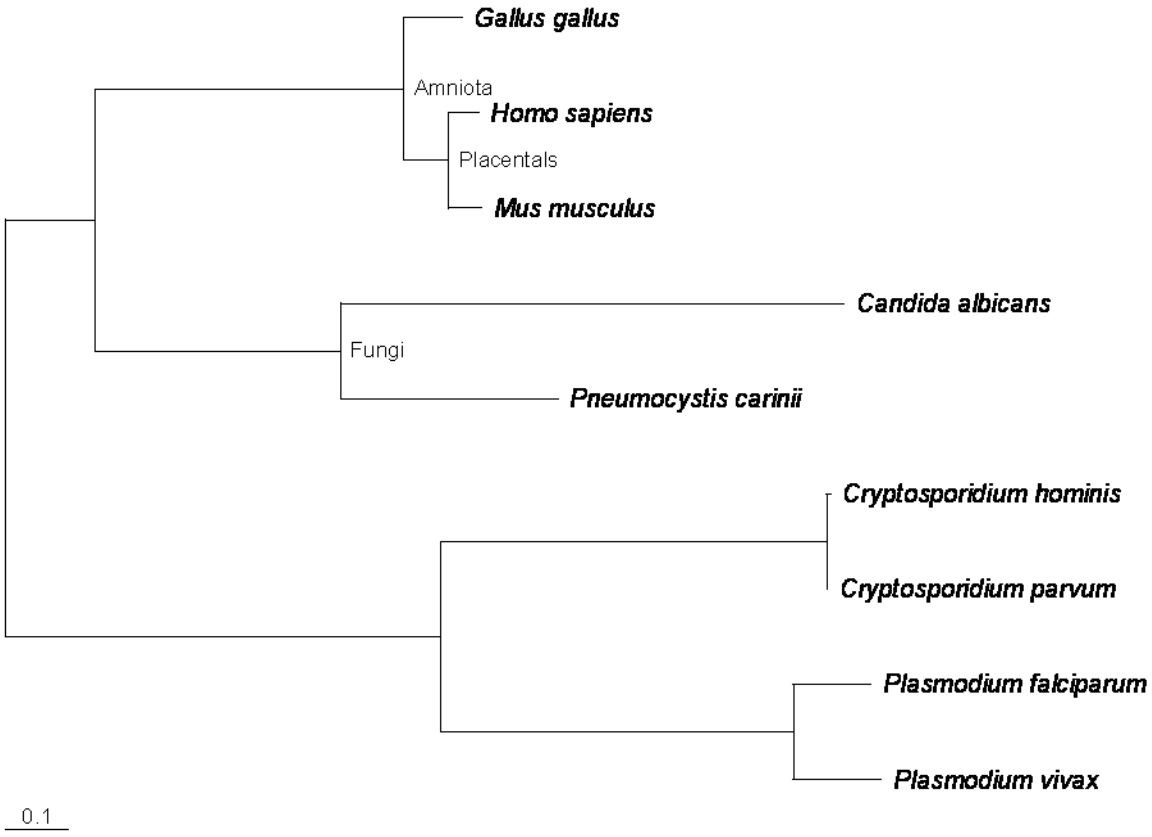


Figure 3.1: Phylogenetic positions of DHFR structures

[100] With regard to searches versus sequence databases, an additional reason to seek to minimize gaps (by increasing the gap opening penalty) was to avoid highly erroneous sequences (due to, for instance, inaccurate intron start/stop signal analysis (Chang, M S S & Benner 2004)).

Among fungi and metazoa, DHFR structures are present for (see Figure 3.1, on page 52):

- *Candida albicans* - a fungus (Ascomycota), of the yeast form
- *Gallus gallus* - chicken
- *Homo sapiens* - human
- *Mus musculus* - mouse
- *Pneumocystis carinii* - a fungus (Ascomycota), of the yeast form

Due to the greater quantity of data (and interpretations of data) present for *Homo sapiens*[101] and (relatively) closely related species such as *Mus musculus*, it was chosen to start[102] with these structures as one end of the chain of structures to be produced (see Figure 3.4, on page 149). *G. gallus* was then used to assist in creating the Uramniota (ancestral amniote) model.

[101] The greater quantity of data (and interpretations of data) includes prior phylogenetic work (Brown, W M *et al.* 1982; Easteal 1990; Gibbs, Collard, & Wood 2000; Glazko & Nei 2003; Kishino & Hasegawa 1989; O'hUigin *et al.* 2002). Said greater quantity of data (and disputes as to its interpretation) can be attributed, at least partially, to the species of the researches and resultant (quite rational) “bias” in the direction of research.

[102] The intent was that the more difficult portion of the work would come after greater experience with the methodology was gained.

The intended end goal was selected to be *C. albicans*. This decision was despite the worrisome presence of a change of nuclear genetic codes - CTG/CUG decoding as serine instead of leucine (Sugita & Nakase 1999a, 1999b) - for *C. albicans* and a few (probably related - see "Tree results", on page 201) other Ascomycota (Fitzpatrick *et al.* 2006; Massey *et al.* 2003), because *P. carinii*:

- Is not well-defined as a species (Stringer 1996), including up to 42% amino acid sequence difference in DHFR between *P. carinii* isolated from the lungs of different species (Ma *et al.* 2001); and
- Had (and has) relatively few sequences (in comparison to *C. albicans*) in the database, especially at the time of the initiation of the research[103], partially due to difficulties in culturing it (Merali *et al.* 1999).

However, it has ultimately not been possible to reach either the end goal of a *C. albicans* structure, or the intermediate goal of the (slightly genetically closer to the fungi/metazoa common ancestor according to current data) *P. carinii* structure (for the sequence corresponding to the deposited *P. carinii* structures), due to time constraints[104]. Further research is planned.

2. Determine sources for phylogenetic sequence data

Since the criteria for the other sequence sources were less strict than the criteria for the central protein, it was found necessary first to come up with a listing of

[103] There are currently (according to a 1/11/08 search on Entrez Genome Projects (http://www.ncbi.nlm.nih.gov/sites/entrez)) 3 genome projects on *P. carinii*, including ones on *P. jirovecii* (from human) and *P. murina* (from mouse), but none are complete as of the writing of this dissertation.

[104] Said time constraints led to attempting to do too many models in a short time frame. In turn, this led to inadequate error correction, causing errors to accumulate - see "8. Examination of models", on page 352.

proteins of interest and then evaluate them for possible suitability. The proteins were those of (local or personal) biological interest, partially to maximize the information available about them and partially to increase the subsequent (for other projects) utility of data gathered/determined about them (e.g., alignments).

Database of structures and species

The first stage of this work involved the creation of a database of structures and the corresponding species from which the structural sequences were derived. This step was made necessary by several factors:

1. PDB files are generally deposited by biochemists[105], not biologists (much less taxonomists/phylogeneticists). Given that the interest of the depositor is likely to be in the structure itself, not in the species from which it came (with obvious exceptions such as *Homo sapiens*), some degree of inaccuracy or confusion is unsurprising. Perhaps the most notable example of this is 1KLK (Cody *et al.* 2000), a DHFR structure noted in the PDB file header as being from *Rattus norvegicus* (the laboratory rat). In actuality, while the DHFR in question was indeed from inside the lungs of a rat, as both the accompanying paper and the sequence (Wang, Y H *et al.* 2001) make clear, it was from *Pneumocystis carinii* growing inside said lungs[106]. More common problems include, especially in older PDB files, the usage of

[105] Please note that this is not meant to disparage the determiners and depositors of structures. As is appropriate given our research's extensive usage of structures (and our insistence on only using structures that have been properly deposited - see "Choice and Availability of Programs and Data", on page 43), we are grateful to those who have done the work to make structural data available to the scientific community.

[106] *P. carinii* is difficult to grow outside of the lungs of another organism (Merali *et al.* 1999; Stringer 1996).

common names instead of scientific names (e.g., "human" instead of *Homo sapiens*).

2. Name changes of species have taken place since the time that many PDB files were produced, particularly for some microorganisms (e.g., various *Burkholderia* species were originally described as *Pseudomonas* species).
3. Existing databases for PDB file species attributions do not include information about cases in which more than one species is known to have the sequence in question. Perhaps the most important case of this for the present work is 1SEJ, which has a DHFR/TS sequence found in both *Cryptosporidium hominis* and *Cryptosporidium parvum*. Another instance of this, important for phylogenetic tree determination, is that the sequence for UBC E2 from 1FBV (chain C) and 1C4Z (chain D) is found in both humans and *Mus musculus* (mouse).
4. Some existing information is confusing, especially to a computer program, in that it uses species names that contain the names of other species (for endosymbionts such as *Wolbachia spp*. and viruses, primarily). This situation increases the likelihood of any "naive" automated means misclassifying the species of origin of the structural sequences in question.

Given the above considerations and others, a database was created of PDB file chains versus the species the protein sequences in question are found in. This database was initially generated using a combination of the PDB file headers and the:

- SCOP (Hubbard *et al.* 1999; Lo Conte *et al.* 2000; Murzin *et al.* 2000),

- SWISS-PROT (Boeckmann *et al.* 2003), and
- PDBeast (Bryant 2004; Wang, Y *et al.* 2002)

databases. Inconsistencies between these resolved either automatically (generally, only when more than one of them indicated the same species) or manually (in favor of whatever organism the ATOM (coordinate) record sequences were found to be from according to a `blastp` search - see page 63). The resulting database, which is available for other work (see "2. Determine sources for phylogenetic sequences" on page 191), contains 37388 chains (from 18583 PDB files) cross-referenced to 1124 species[107]. This database was used to narrow down the possibilities, in favor of those with multiple, significantly divergent (less than 65% identical) sequences with known (3D) structures.

[107] As can be seen from these numbers, the database is rather larger than strictly necessary for the present work, although the manually-edited portion of it does focus on proteins potentially of use for this work (said manually-edited portion does not include viral-derived sequences, for instance, except in listing them in a file of structures unsuitable for this or other reasons).

Other proteins used

The (non-DHFR) proteins[108] ultimately used were (in alphabetical order):

- Alcohol dehydrogenase class 1 (ADH1; E.C. 1.1.1.1). Note that three isozymes (alpha, beta, and gamma) of this protein are found in primates (Buhler *et al.* 1984; Cheunq *et al.* 1999) To take this into account, the sequences for each were placed in a separate "partition" in MrBayes, while other sequences (including ADH1 from non-primates) were duplicated[109] three times.
- Catalase (E.C. 1.11.1.6)
- Cellulase A (glycosyl hydrolase 5; exo-1,3-beta-glucanase; E.C. 3.2.1.58) (Coutinho, P M & Henrissat 1999; Coutinho, Pedro M & Henrissat 2007; Henrissat 1991; Henrissat & Bairoch 1993; Henrissat *et al.* 1995; Henrissat & Davies 2000)
- Cinnamyl alcohol dehydrogenase (Cinnamyl ADH; E.C. 1.1.1.195). Note that the identification of 1UUF (YAHK_ECOLI) as a cinnamyl alcohol dehydrogenase is not based on sequence (nor on the structure), but on enzymatic activity assays performed locally (Chase 2005; Khalid 2001). The determination of the structure was as a part of a structural genomics

[108] Note that, in much of the program code and datafiles, homologous proteins are labeled as being in "groups" (e.g., the ADH1 group). Also note that, for many of the above, the E.C. should be looked up (IUBMB 1992) to provide further information if desired.

[109] This procedure was used so that ADH1 was not weighted more than other sequences; unfortunately, MrBayes lacks other weighting methods, and it was concluded that it was not worth the programming (including debugging) effort involved in adding such capabilities.

(Goldsmith-Fischman & Honig 2003) project on *E. coli*, without the authors (Aberqel *et al.* 2003) being sure of the function[110].

- Copper/Zinc-Containing Superoxide Dismutase (Copper/Zinc SOD or CuZnSOD; E.C. 1.15.1.1)
- Eukaryotic initiation factor 2a (eIF2a). Note that initiation/elongation factors previously used for phylogenetics and found to be problematic[111] were avoided (Gaucher, Miyamoto, & Benner 2001; Lopez, Forterre, & Philippe 1999).
- Eukaryotic initiation factor 4e (eIF4e)
- Eukaryotic initiation factor 6 (eIF6). Note that, despite the name, a structure of this protein is known (1G61) from an archaeon*, Methanocaldococcus (Methanococcus) jannaschii*.
- Eukaryotic termination factor 2a (eTF2a)
- Glutathione-S-Transferase (GST) Class Pi (E.C. 2.5.1.18)
- Glutathione-S-Transferase (GST) Class Sigma (E.C. 2.5.1.18) and Glutathione-requiring Prostaglandin D Synthase (E.C. 5.3.99.2); this enzyme varies in function, albeit not in its usage of glutathione, between vertebrates and invertebrates (Jowsey *et al.* 2001);
- Glutathione-S-Transferase (GST) Class Zeta (E.C. 2.5.1.18)

---

[110] They thought it likely to be a (zinc-dependent) ADH, but were not sure of this; nor did they identify it as not acting to any significant degree (K_m over 1M) on ethyl alcohol (Chase 2005; Khalid 2001).

[111] The problems were due to covarion (different rates for different residues on different parts of the tree), long branch attraction (see footnote 52 on page 27), or other effects. It was unfortunately not possible to use MrBayes' covarion setting to try to compensate for the first of these (see page 99, footnote 200).

- Hemoglobin V/Alpha; hemoglobin is found in a tetramer of 2 alpha and 2 beta chains in most vertebrates. It is found simultaneously in multiple types ("isohemoglobins") in fish, and in jawless fish (e.g., lampreys) has a variable degree of association, sometimes being monomeric and when multimerized not resembling the alpha/beta interface. Lamprey (and hagfish) globin appears to have descended from a common ancestor of alpha and beta globins[112] (Mito *et al.* 2002; de Souza & Bonilla-Rodriguez 2007). Hemoglobin V, as identified by sequence similarity, is the most common component in two different lamprey species, *Petromyzon marinus* and *Lampetra fluviatilis* (Hombrados *et al.* 1983).
- Myoglobin (Dutheil & Galtier 2007; Neher 1994)
- Orotidine-5'-phosphate decarboxylase (ORO), as described above.
- Poly(A) Polymerase (E.C. 2.7.7.19)
- RecA (also known as Rad51 for eukaryotes and RadA for archaea). This enzyme is involved in DNA recombination and repair (Sandler *et al.* 1996; Sung *et al.* 2003; Thompson, F L *et al.* 2005)[113].
- Sorbitol Dehydrogenase (Sorbitol DH; E.C. 1.1.1.14)

---

[112] That alpha and beta globins arose via gene duplication is additionally confirmed by the arrangement of the alpha and beta globin genes in teleost fish, namely directly adjacent (de Souza & Bonilla-Rodriguez 2007).

[113] This protein's recombination role was considered particularly valuable in case of (undetected) horizontal gene transfer events, since to the degree that such events are important, the proteins involved would be likewise important. In other words, if one considers sufficient genetic migration to create a new species (including through said migration providing evidence that two existing "species" were/are not actually species), then the proteins enabling said genetic migration are determining species identity by their function. Analogously, proteins involved in maintaining the gene-flow separation between species (e.g., pheromones, sperm/egg proteins, or mating behavior determinants) would be of interest. However, these tend, by their very nature, to mutate at a fast enough rate (at least one should mutate significantly with each speciation event) to make their rate of change too high for a wide-ranging phylogenetic study (similarly to DNA sequences as compared with protein sequences).

- TATA-Binding Protein (Johnson, S A S *et al.* 2003) - abbreviated TBP, and also known as TF2D
- Triosephosphate Isomerase (TPIS; E.C. 5.3.1.1)
- Ubiquitin conjugating enzymes of the E2 family (UBC; E.C. 6.3.2.19); the aligned portion is the conserved UBC domain (Ardley *et al.* 2000; Iyer, Burroughs, & Aravind 2006; Winn *et al.* 2004), according to InterPro (Mulder *et al.* 2007) - see http://www.ebi.ac.uk/interpro/ISpy?ac=P21734 and http://www.ebi.ac.uk/interpro/ISpy?ac=P68036.

Some additional proteins (see "Appendix F: Proteins removed", on page 373) were used for tree determination for several stages. However, when the stage was reached for finding the fungi/metazoa ancestral DHFR sequence, an attempt was made to align in the *Neurospora crassa* DHFR sequence (to assist in finding said sequence) and it was realized that the *Neurospora crassa* DHFR sequence was not reliable (it contains large Gln (Q) insertions). The removal of this species (from among those considered to have a (usable) DHFR sequence) eliminated the usefulness[114] of these proteins for further work on[115] the DHFR ancestral sequence determination.

Structures and sequences

See "Appendix A: PDB files/chains used", on page 366, for information on the PDB files used for this and the prior step; see "Appendix B: Important PDB files/chains used", on page 367, for which PDB files were of most importance and

[114] Please see "Appendix F: Proteins removed", on page 373 for more details.
[115] However, they are retained in the alignment database for future usage.

for more information on the quality of the structures. As with DHFR, PDB files were required to have backbones and side chains (not be alpha carbon only) and have an estimated[116] RMS error (uncertainty) of 0.6 Ang. or less (with 0.6 Ang. allowed only for NMR). For X-ray crystallographic structures (which were preferred to NMR), the resolution was required to be 3.3 Ang. or less. No electron microscope or similar structures were used. When possible, only structures with estimated RMS errors of 0.25 Ang. or less and resolutions of 2.5 Ang. or less were used. If alternate coordinates were present in X-ray crystallographic files, then the highest-occupancy alternate coordinates were chosen (or coordinates labeled "A" if tied).

Following the location of the initial PDB files for each of the above, further PDB files were located using BLAST versus the sequences found in the PDB file ATOM records (see page 51), with any initiating methionine, His tags, *etc.* removed. This BLAST stage used NCBI's `blastp` with default settings (e.g., BLOSUM62) aside from the gap-opening penalty increased (from the default 11) by 1 to the maximum, 12 (since gaps are known to be problematic with alignments, we sought to reduce the number of files found with significant gaps).

[116] The RMS error was estimated via the generation of points on the Luzzati plot (Luzzati 1952, 1953) followed by the interpolation (using Math::Interpolate's robust_interpolate function (Zajac 1999)) of the RMS. This procedure was necessary because the Luzzati equation gives the R-factor from the resolution and RMS, not the RMS error from the resolution and R-factor around; no closed-form solution to deriving the RMS appears to be available. Free R-factors were used instead of R-factors when available, with structures lacking Free R-values penalized; see "Appendix B: Important PDB files/chains used" on page 367. NMR structures were treated as having a resolution of 2.5 Ang. and a Free R-value (R-factor) equal to the worst encountered among actual crystallographic structures to derive an RMS, although an examination of the divergence among ensemble models is recommended for future work.

PDB files less than 30% identical to others[117] were excluded, since the "twilight zone" of structural alignment is thought to start below (approximately) this level (Rost 1999; Yang, A-S & Honig 2000b).

The ATOM-record derived sequences from all PDB files considered usable were then queried versus the NCBI nr database (Wheeler *et al.* 2000) in a further BLAST search. Again, NCBI's `blastp` was used, but with BLOSUM80[118] (since the desired sequences were closer, with at least 65% identity; see item F, on page 20). The results of these searches were saved and interpreted automatically by "interpret.protein.files.pl"; the supplemental file "Makefile.prior.txt"[119] has more information on the files used/required. The results can be examined in supplemental files[120] "interpret.protein.files.txt.new.txt" (with all proteins) and "important.protein.files.all.txt.new.txt" (with only proteins from species considered promising for further work, based on the number of sequences from them and whether they had a DHFR or DHFR/TS sequence).

With regard to whether sequences were 65%+ identical, the percent identity from the `blastp` search was not directly used, since it is only the percent identity over the *aligned* part of the sequences. Instead, the number of identical residues

[117] Others, that is, outside of that 65%+ identical cluster - in other words, except to others that were more than 65% identical, i.e., that did not need structural alignment in the first place.
[118] The defaults were also changed to increase the gap-opening penalty. The penalty was increased by 1, to 11 (the default for BLOSUM80 is 10, and the maximum is 11); again, this was done to minimize the number of sequences used with significant gaps.
[119] This file is used by the program `make` (Stallman, McGrath, & Smith 1998) to direct the first part of the sequence processing.
[120] These are also available online at http://cesario.rutgers.edu/easmith/research/proteins/interpret.protein.files.txt.new and http://cesario.rutgers.edu/easmith/research/proteins/important.protein.files.all.txt.new,

(modified for gaps - see below) was compared with 0.65 times the number of identical residues in the best[121] comparison in the `blastp` output file in question[122]. Given the problematic nature of gaps in alignments and phylogenetics, if the number of gaps was over 1% of the highest number of identical residues in the output file (minimum 1), the effective number of identical residues was decreased by the number of gaps minus the number of allowed gaps. Please see "interpret.protein.files.pl" for the exact details if desired.

Usage of polymorphism

One question in using sequence search results is what to do when more than one sequence shows up from a species; a related question is what, if anything, to do with variants noted in, e.g., SWISS-PROT's VAR records (Boeckmann *et al.* 2003; UniProt 2005). While at first it might appear that simply ignoring all but one record would be best, there are reasons for not doing so:

- As noted above, a mechanism for handling polymorphism can also be useful for some isozymes;
- In some cases, the record chosen would be arbitrary, because more than one sequence was equally close to the others;
- In some cases, "polymorphism" was due to uncertainty about species identifications (see "Resolution of species ambiguities", on page 77), making the removal of such particularly undesirable;

---

respectively.

[121] Best, in this context, is the sequence with the highest number of identical residues.

[122] This procedure was done because of differences in some cases between local and NCBI's definitions of sequences (due to the use of ATOM records or otherwise) meaning that no

- As discussed below under "Partitions: Gamma, Invariant, Rate" (on page 105), a significant number of different sequences are helpful in determining what locations in proteins are variable, and what amino acid substitutions are acceptable (Pollock & Bruno 2000). Logically, if, for instance, variant proteins exist in healthy members of a species (as in, species members who do not show any health problems due to the variants), then the substitution is an acceptable one, and the ancestral protein may well have had said different amino acid.

Therefore, the decision was made to make use of variations.

Criteria for polymorphic sequences used

Variations in two categories were used:

1. If the initial (`blastp`) sequence search:
   - turned up more than one candidate protein sequence from a given species (including from combined species), and
   - an initial manual examination did not indicate any obvious problems[123];

   then the sequence IDs were entered into the datafiles[124] "proteins.polymorphism.manual.txt" or (for ADH1 Alpha/Beta/Gamma) "split.polymorphism.txt".
2. VAR records were present in SWISS-PROT and the variant in question appeared, on manual review, to be overall neutral or beneficial[125].

---

sequence was 100% identical.

[123] Examples of problems would include a fragment, pseudogene, or association with a disease phenotype sans known heterozygote advantage - e.g., amyotrophic lateral sclerosis type 1 for CuZnSOD (Wroe & Al-Chalabi 2007).

The datafiles "proteins.polymorphism.manual.txt" and "split.polymorphism.txt" were processed by the program "list.polymorphism.pl" into "list.polymorphism.txt"[126], which was used for subsequent processing for both of the above categories. For the second, SWISS-PROT (both the above-referenced version and that as of May of 2006, to allow for variants added after the Feb 2005 version) and the data file "list.polymorphism.txt" were processed by the program[127] "extract.sptrembl.polymorphism.pl" to extract the variation information. The resulting file[128] was put together with "list.polymorphism.txt" and material from the earlier alignments (see "Multiple alignments", on page 93) by the program "align.polymorphism.pl". As well as aligning (by the same mechanism as "integrate.structural.align.1.pl" in regard to "compromise" alignments - see below) in the variant sequences, this program evaluated the variants versus the main sequence, and the main sequence versus sequences

---

[124] These are available as supplemental files and under http://cesario.rutgers.edu/easmith/research/proteins/.

[125] One heterozygote-advantage mutation group, namely those of Hemoglobin Alpha producing alpha(+)-thalassemia (please see the SWISS-PROT (Boeckmann *et al.* 2003; UniProt 2005) entry for HBA_HUMAN and the material referenced therein), was of concern in this regard. Alpha(+)-thalassemia can be produced by either:

- the lack of any function in two or more alpha-globin genes (there are normally two on each chromosome, or four in all); or
- most/all of these genes being semi-functional.

Mutations resulting in a nonfunctional hemoglobin product were discarded (as essentially being at least as bad as lacking the hemoglobin gene entirely, with deleted genes not being applicable to our study). Likewise, variants resulting in a lower level of functionality were discarded unless both:

A. at least three of the 4 genes being mutated were required to cause actual disease; and
B. evidence was present for alpha(+)-thalassemia's resistance to (severe) malaria in heterozygotes being present for the mutation in question.

[126] This file is available as a supplemental file and under http://cesario.rutgers.edu/easmith/research/proteins/.

[127] Instructions to the program on which variants to skip are in the data file "list.polymorphism.txt" (for which sequence IDs to attempt VAR record extraction on) and in the program code.

[128] The file is “extract.sptrembl.polymorphism.txt"; it is available as a supplemental file and under http://cesario.rutgers.edu/easmith/research/proteins/.

from other species - in that cluster (65%+ identical sequences) if possible, versus all other sequences if not - by the following criteria:

1. What the percent identity was of the aligned residues;
2. What proportion (of the maximum possible) of the residues were aligned (variants with gaps versus the main sequence were penalized); and
3. What the similarity (as judged by the locally created matrix ESIMILARITY - see "Appendix G: ESIMILARITY matrix", on page 374) of the amino acids was.

All sequences derived from SWISS-PROT VAR records, and all sequences that appeared to be derived from species ambiguities, were added, as were any other sequences that were better on the above criteria than any "bad" sequence (one from another species). If the cluster did not have at least 30 sequences, and there were other possible sequences to add, they were added, in order of closeness, until there were at least 30 sequences in the cluster.

One error was found at this point in that some sequences from chains had not been included properly; this was corrected by "align.polymorphism.add.seqs.pl", outputting the file "align.polymorphism.add.seqs.txt". When it was considered desirable to add the DHFR sequences to the phylogenetic dataset[129], these were added manually to this file, after conversion of the alignment from Stockholm

[129] DHFR sequences were added into the alignment after all tree rearrangements except two were evaluated (see "Tree results", on page 201); the exceptions were:

1. The placement of Rodentia at the root of placental mammals, instead of the earlier arrangement with Primates - see “Tree search with Mammalia (subset)” on page 316;
2. The placement of *P. carinii* and *S. pombe* as grouped together at the base of Ascomycota (the group containing most studied yeasts), as opposed to the earlier arrangement of first

format (see “5. Alignment of central sequences”, on page 336) to aligned FASTA format by the HMMER-associated program "sreformat" (Eddy & Birney 2003).

Creation of “full” species

The method chosen[130] to input the polymorphic sequences into MrBayes was to expand single species ("real" species) into multiple "full" species[131]. This task was performed by the program "combine.structural.align.groups.pl". This program, if there was polymorphism, called an external program, "consensus5.multiple.pl", to determine which of the sequences was closest to a consensus[132]. If this program was unsuccessful in this, then it would call another program, "consensus4.multiple.pl", that could create an ambiguity-coded sequence if need be. Such an occurrence was, however, avoided if possible by "combine.structural.align.groups.pl"[133], by passing information to "consensus5.multiple.pl" as to which sequence(s) were the most suitable (e.g.,

---

S. pombe then *P. carinii* branching off - see “Tree rearrangement for *P. carinii*, *S. pombe*” on page 320.

[130] This method was chosen instead of using NEXUS polymorphism coding (with alternative amino acids in brackets or parentheses) for two primary reasons:

- Using NEXUS polymorphism coding would have lost information as to what amino acid was most common (especially in cases when there are multiple sites for polymorphism such as with hemoglobin and thus most variants may be the same in many locations);
- The difficulties with polymorphism and/or ambiguity (Huelsenbeck *et al.* 2006) in MrBayes (despite some code alterations - see item 2 under “MrBayes code alterations”, on page 98).

Some compromises had to be made on this (see “Species, polymorphism reduction” on page 70; also note on page 68 regarding the consensus/compromise sequence).

[131] Note that these have been collapsed into the real species in most of the tree results (see "Tree results", on page 201), by either the program "tree.simplify.full.pl" or the program "create.tree.section.pl". Trees containing information on the reliability of nodes from tree searches (see “Tree searches”, on page 299) are exceptions; both of these programs would have removed that information, since it is coded as numeric “names” for internal nodes (the removal of said information is a limit of these programs that time limits have not permitted removing).

[132] Closeness was judged by identity then by similarity (as per "Appendix G: ESIMILARITY matrix", on page 374).

[133] It was also avoided by later reductions in ambiguity coding that took into account sequences from other species (see “Further sequence processing: Ambiguity-coded polymorphism reduction”, on page 94).

base sequences were preferred to sequences from SWISS-PROT's VAR records). The resulting sequence was called the "compromise" sequence, and the original sequence matching it (if any) was classified as among the "best" sequences[134]. The "compromise" sequence could be altered to become less ambiguous in some cases, if it was necessary to decrease the degree of polymorphism of some species/proteins[135]. A set of "full" species was then created, in which most of the sequences were the "compromise" sequence, but one group had one of the other sequences. This task was done by going through each of the[136] real (non-"compromise") sequences in turn, having the sequence chosen as the sequence used for its corresponding protein (group), and having the other protein sequences be the "compromise" sequences for those proteins[137] (groups). These "full species" were then sorted for quality. This sorting was based on criteria such as what sequences had been decided to be the "best", how close the sequence (in terms of identity and similarity) to these was, whether it was the "compromise" sequence, whether the "compromise" sequence had been altered, *etc.*. The aim was to have the most representative set of sequences being the first one (signified as, e.g., "*Homo_sapiens.01*" or

[134] Some "best" sequences were also manually specified (please see the program code) to make sure that, for instance, DHFR sequences with "polymorphism" due to alternate alignments (see "Alignment using HMM", on page 129) were considered good.
[135] See "Species, polymorphism reduction", on page 70; amino acids found only in the "main" or "best" sequence(s) and not in those being removed would be removed from the ambiguity coding, since the "main"/"best" sequence(s) would be retained.
[136] If polymorphism had been reduced, then only the remaining sequences would be used for this.
[137] In other words, if there were 2 proteins (groups), and 3 sequences for the first protein (e.g., "A", "S", "T"), each of these would be paired with the "compromise" sequence for the other protein (group) to create one set of "species". Each sequence for the second protein (group) would then be paired with the "compromise" sequence for the first group (e.g., "S").

"*Arabidopsis_thaliana.1*"), with the rest having higher numbers[138] (e.g., "*Homo_sapiens.02*" or "*Arabidopsis_thaliana.2*").

The "full" species were constrained to be together in MrBayes for any tree rearrangements, with an increasing distance from their common root as the number of the "full" species became higher (e.g., "*Homo_sapiens.01*" would branch off before "*Homo_sapiens.02*", which would branch off before "*Homo_sapiens.03*"). These constraints were added, along with ones for species groups (see "Appendix I: Species groupings used", on page 376), by the program "nexus.add.kingdom.constraints.pl".

Species, polymorphism reduction

Unfortunately, it was necessary to eliminate some of the polymorphism and species (although they are still present in the database prior to processing by "combine.structural.align.groups.pl"), due to the number of effective species that would otherwise be generated by the above[139] and resultant overload of data[140]. This removal took place within "combine.structural.align.groups.pl" (after data on amino acid frequencies had been extracted for later usage - see "Partitions: State

[138] For examples of trees with "full" species included, please see under "Tree searches" (on page 299). Other tree displays use the distance to the "full" species closest to the root (this is usually, but not always, the lowest-numbered "full" species).

[139] Another reason this was necessary was the presence of some species with very little sequence data, which appeared to be problematic for:

- branch lengths (see "Tree distances", on page 113); and
- placement in the phylogeny (according to "compare.trees.problems.pl") for the initial (done prior to this reduction) tree searches.

[140] This may be described as piloting between the Scylla of too many sequences to handle and the Charybdis of too much loss of data.

frequencies", on page 107); both some species and some polymorphism were removed. The reduction in species used four primary criteria:

1. Whether the species had a known DHFR (or DHFR/TS) sequence - the removal of such species was avoided;
2. How many different proteins were known for the species - in particular, species with only one protein known were considered for removal;
3. How many amino acids were in the proteins known for the species - species that fell below limits[141] established by examining which species were being inconsistent between tree "runs" for branch lengths (see under "Tree distances", item 19, on page 123) - or, using earlier tree searches, between search results from different runs (as judged by “compare.trees.problems.pl") - were considered for removal;
4. Whether the removal of the species would result in a cluster being reduced below 20 or 30 sequences (see "Partitions: Gamma, Invariant, Rate", on page 105) - if so, its removal was avoided.

Similarly, the level of polymorphism in proteins other than DHFR was reduced if it was possible to do so without going below 20 or 30 sequences in the cluster in question[142], with priorities in this based on, for instance, whether the species in question had a DHFR (or DHFR/TS) sequence known.

---

[141] If the species in question was bacterial, then the limits used were stricter (required a higher number of amino acids). This was due to the number of bacterial species in the database relative to the concentration of the present work on eukaryota and the difficulties in creating a reasonable starting tree for the bacterial species - due to, e.g., horizontal gene transfer (Gogarten, Doolittle, & Lawrence 2002; Gogarten & Townsend 2005). Also kept in mind for archaea and eukaryota were the results from Tree-Puzzle, in which it appeared that RecA (RadA, for Archaea) had sufficient information content to be usable for tree determination even for species with only its sequence known.

[142] This was particularly so if it was currently judged possible to use gamma rate variation (again,

3a. Creation of a rough starting tree

Initial sources

The program Tree-Puzzle (von Haeseler & Strimmer 2003; Schmidt *et al.* 2002; Strimmer & von Haeseler 1996; Strimmer, Goldman, & von Haeseler 1997; Strimmer & von Haeseler 1999) was initially used for tree creation, but encountered problems due to missing data as noted earlier (see footnote 53, on page 30). It was, however, successful in creating a reasonable tree[143] for Archaea; the tree from it was used for Archaea in subsequent work, with the addition of *Archaeoglobus fulgidus* with other Euryarchaeota. The initial basis of the rest of the starting tree was primarily the NCBI taxonomy database[144] (Wheeler *et al.* 2000) as of August 21st, 2004 (Bischoff *et al.* 2004), with interpretation and the needed subset extracted by "nexus.create.ncbi.tree.pl" (from the processed version - see “Appendix D: NCBI taxids and alternate species names”, on page 370). This taxonomy is not, however, sufficiently non-polytomous (see footnote 208, on page 101). Moreover, a usage only of it as a basis for starting may be create a bias toward a more "classical" phylogeny in some cases than may actually be argued for by the evidence[145]. Accordingly, a need was felt to "blur" the taxonomy in question further, by combining it using

---

see "Partitions: Gamma, Invariant, Rate", on page 105) on the protein section in question - if so, it would be particularly undesirable to reduce the number of sequences below 20.

[143] All quartets could be assembled into a tree congruent with the NCBI taxonomy (see above).

[144] See http://www.ncbi.nlm.nih.gov/Taxonomy/ for the latest version of this database.

[145] Since large, frequent changes in the NCBI taxonomy would be likely to be problematic, it appears that they typically wait until some measure of agreement has happened in a field (a process that, at the worst, may involve a generational change, particularly if philosophical differences are the problem) prior to changing the taxonomy.

quartets (see "Usage of quartets", on page 74) with a taxonomy created using genetic evidence not otherwise usable by this research's methodology. This taxonomy, which was exclusively of Eukaryota except for Bacteria/Archaea as a root, mainly used nuclear genetic code changes, gene splits and fusions (Stechmann & Cavalier-Smith 2002, 2003), and some heuristics to represent generally agreed-upon groupings (a subset of those found in "Appendix I: Species groupings used", on page 376); these characteristics were interpreted by the program "determine.parsimony.species.pl". The output was then interpreted via the PHYLIP (Felsenstein 1993) parsimony[146] program PENNY using heuristically determined weights[147]; the results were then put together by the PHYLIP program CONSENSE to yield a majority rule consensus tree. The resulting tree cannot be described as particularly realistic. It was, however, at least successful in "blurring" the NCBI taxonomy (see "Usage of quartets", on page 74) with what can perhaps best be described as an "interesting" set of phylogenetic hypotheses. This ("parsimony") tree can be found in supplemental file[148] "MyTree0001.nexus.txt"; please note that the author *strongly* advises against using it, in general, for any purposes other than ones similar to those of the present research.

The next stage was to gather results from prior phylogenetic studies. The primary source for trees containing this information was TreeBASE (Sanderson *et al.*

[146] Parsimony was used because there is not an adequate amount of evidence of the likelihoods of, for instance, nuclear genetic code changes.
[147] See the supplemental file “weights.single.txt”, also available via http://cesario.rutgers.edu/easmith/research/trees/weights.single.txt.
[148] It is also available at http://cesario.rutgers.edu/easmith/research/trees/MyTree0001.nexus.

1993); a listing of the TreeBASE trees (and other sources used, such as when TreeBASE lacked usable trees giving information about the species arrangement in question) is in "Appendix C: Other sources for initial tree", on page 369. It was necessary to reformat the material from TreeBASE and translate it into a common set of species. This included the clarification of any species ambiguities (see "Resolution of species ambiguities", on page 77) and splitting any above-species names into their component species (according to the modified NCBI taxonomy - see "Appendix D: NCBI taxids and alternate species names", on page 370). This process was done by "nexus.interpret.treebase.trees.pl" and, when the need to split up the problem was later realized (see below), “nexus.add.groups.2.pl”[149].

Usage of quartets

The NCBI taxonomic and "parsimony" trees were initially split into quartets (see "Tree construction methods", on page 28) by the program "nexus.find.init.quartets.pl"; quartets that were contradictory between the two trees were noted as "dual" and as having both species arrangements as possibilities. It was realized that the number of quartets this process produced was too many to deal with in later work. The trees were therefore split up into groupings (similar to those in "Appendix I: Species groupings used", on page 376), with the splitting[150] of the NCBI tree being via multiple programs[151]. The

[149] See “nexus.add.groups.2.txt”, in the supplemental file “trees.tar” (in UNIX “tar’ format) or at http://cesario.rutgers.edu/easmith/research/trees/nexus.add.groups.2.txt, for the output of the latter.

[150] In this context, "splitting" means the substitution of the names of larger phylogenetic groups for those of multiple species inside the respective groups.

splitting for the "parsimony" tree was done by “nexus.create.ncbi.subtrees.MyTree0001.pl”. The extraction of quartets was then done via "test.find.quartets.1.pl"; metazoa quartets were then split out manually, and non-metazoa quartets were split into groups via "nexus.split.non_metazoa.quartets.pl". An attempt was then made to clean up the quartets generated, including creating any new quartets implicitly present and removing any that were contradictory (Piaggio-Talice, Burleigh, & Eulenstein 2004; Willson 2001), using the program "nexus.cleanup.quartets.pl". The implementation of this method was based on the program Rectify from Quartet Suite (Piaggio-Talice & Piaggio 2003), but with the modification that, instead of the total number of contradictory quartets to a particular arrangement of species being counted to determine whether a new quartet could be generated, only quartets with no contradictions among already-known quartets were generated. It unfortunately appears that the results were not as intended. A number of instances of quartets were seen that contradicted one or the other of the NCBI or "parsimony" trees. Fixing this problem required a considerable amount of manual intervention to correct (discussed on page 76).

---

[151] These programs (see “Appendix P: Perl programs created”, on page 415) include nexus.create.ncbi.subtrees.pl, nexus.create.ncbi.subtrees.2.pl, nexus.create.ncbi.subtrees.3.pl, nexus.create.ncbi.subtrees.4.pl, nexus.create.ncbi.subtrees.5.pl, nexus.create.ncbi.subtrees.6.pl, nexus.create.ncbi.subtrees.7.pl, nexus.create.ncbi.subtrees.8.pl, nexus.create.ncbi.subtrees.9.pl, and nexus.create.ncbi.subtrees.10.pl. Note that some of these programs took the output of earlier programs and split it further, due to realizations at later points in the process of the need, for computational reasons, to split the problem up further.

The TreeBASE trees[152] were then evaluated for consistency[153] with the previously found quartets from the NCBI and "parsimony" trees. The degree of consistency was noted, and when the TreeBASE trees contradicted each other, the most-consistent tree(s) were used to put in the missing quartets that were (implicitly) responsible for polytomies[154]. The evaluation of quality of the TreeBASE trees was performed by several programs[155]. The evaluations generated by these programs were used by “nexus.get.quartets.2.pl” (run by “nexus.get.quartets.2.wrapper.pl” for each kingdom). Due to some data being missing from TreeBASE, along with some problems probably due to the previously-noted (see page 75) difficulties with the "cleanup" program, it was necessary to do some manual intervention in order to put together the ultimate starting tree. This process used the sources noted in "Appendix C: Other sources for initial tree", on page 369, with quartets derived from manually input trees by the program "nexus.find.overall.quartets.1.pl"; the changes in question can be found in the source code of the programs noted above. The final assembly of the subtrees into the final version of the starting tree[156] was done manually.

---

[152] In addition to trees from TreeBASE itself, the tree from one prior paper that was noted as being particularly valuable but missing from TreeBASE (Stechmann & Cavalier-Smith 2003) was put in as "MyTree0002.nexus".

[153] This process was necessary, not only due to disagreements between studies, but due to that some TreeBASE trees are entered as *alternative* trees that were checked and found to be less likely to be correct than other trees, likewise also entered.

[154] For trees from different studies, this was done via weighting; for trees from the same study, the most consistent one with data on the quartet under consideration was used (see footnote 153, on page 75, for why).

[155] These (see “Appendix P: Perl programs created”, on page 415) were: nexus.get.quartets.recover.pl, nexus.get.quartets.recover2.pl, nexus.get.quartets.recover3.pl, nexus.get.quartets.pl, and nexus.get.quartets.kingdom.pl. Some of these were due to realizations of the size of the problem necessitating halting of programs and subsequent recovery of data.

[156] E.g., substituting the “Archaea” subtree for the “Archaea” label in the Eukaryotic subtree once the latter was assembled.

Following the derivation of a set of quartets, they were translated into a tree by, depending on the degree of confidence in the results, either:

- first processing through "quartets.to.wr.modified.pl", then through the program Rectify (Piaggio-Talice & Piaggio 2003) using the "-mf" option, if the quartets were considered dubious (this was done with the "Other Eukaryota", namely those other than Fungi, Metazoa, or Plants/Algae (Viridiplantae)), then through Assemble (Piaggio-Talice & Piaggio 2003) to generate the starting (sub)tree;
- first processing through "quartets.to.weights.pl", then directly through the program Assemble (Piaggio-Talice & Piaggio 2003) to generate the starting (sub)tree.

The subtrees were then assembled together; please see "3a. Creation of a rough starting tree", on page 192, for the results.

Resolution of species ambiguities

Concerning questions as to what species a sequence actually came from, and of species versus subspecies in general, the approach followed was generally that of "lumping" species together when in doubt. This practice was not due to any particular philosophical preference for fewer species in the context of phylogenetic studies (as opposed to, for instance, conservation biology), but:

1. Due to the likelihood of confusion between species/subspecies, especially when sequences (including those used in structural determination) were entered by those not specializing in the taxonomy of the group in question

and/or changes have taken place in the nomenclature used (e.g., *Thermus thermophilus* was put together with *Thermus aquaticus*);

2. Due to evidence of (significant, recent) genetic interchange; as well as such interchange contradicting the biological definition of species, these combinations were done to minimize problems with horizontal gene transfer and resultant confusion between gene trees and species trees. For instance, *Canis lupus* (wolves), *Canis rufus* (red wolves), *Canis familiaris* (dogs), and *Canis latrans* (coyotes) were treated as the same species, *Canis lupus*[157], due to evidence of significant interbreeding (Roy *et al.* 1994).

In such cases, the species name used was not necessarily the one determined as correct[158] by the most authoritative sources available at the time (particularly given that such can change and indeed may be in dispute), but rather whatever species name was most convenient (e.g., was mostly commonly used in the literature and thus likely to be recognized). Please see "Appendix D: NCBI taxids and alternate species names", on page 370, for more detailed information.

3b. Alignment of other sequences

All structural alignments were manually reviewed, either here or elsewhere. Structural alignments were from two categories of sources, namely databases created elsewhere and alignments done locally.

[157] This name was chosen due to the derivation of dogs from wolves and the greater recognizability of *Canis lupus* as opposed to *Canis latrans*.
[158] For instance, the “Imperfect State” name (e.g., *Trichoderma reesei* instead of *Hypocrea jecorina*) was frequently used in the Ascomycota.

Previously created structural alignments

Structural alignments from external sources were derived from 3 databases:

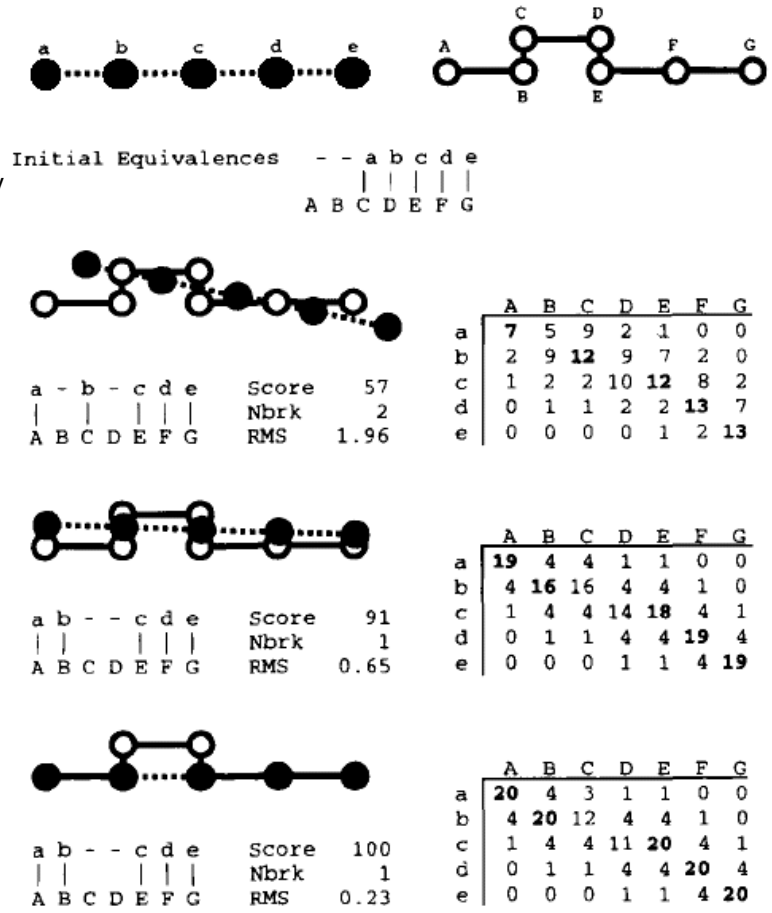
- 3D-ali[159] (Pascarella, Milpetz, & Argos 1996). This database is also referred to as 3D_Ali. It is a database of 3D structures and associated sequences, with the structural alignments based on publications by the depositors of the structures in question. While this database may be considered somewhat out of date, it (and its predecessor from 1992) have been used for other purposes successfully (Vogt, Etzold, & Argos 1995; Wallqvist *et al.* 2000), and its sources are knowledgeable about the structures to be aligned.
- Pfam (Bateman *et al.* 2002): This is the Protein Families database, a collection of protein multiple sequence alignments, with the "seed" alignments (used in this work) being manually reviewed (in some cases, manually created). This database was used only when it appeared that the alignment used structural information, judging by the presence of structural references in the database.
- HOMSTRAD (de Bakker *et al.* 2001; Mizuguchi *et al.* 1998): The Homologous Structure Alignment Database is primarily[160] a collection of structures, aligned initially programmatically (using three different programs) but with the alignments in question reviewed manually.

[159] A copy of this database will be made available locally if necessary; the Argos Group has left EMBL, and it appears that links to the database are no longer being maintained as of the writing of this dissertation.

[160] HOMSTRAD has recently started including some sequences without known (3D) structures in its alignments. These were not used in the present work.

Figure 3.2: Structal method (Gerstein & Levitt 1996, 1998)

1. At **I** are two highly simplified structures (ABCDEFG and abcde) in an arbitrary initial orientation.
2. An initial equivalence is chosen (see main text). Using this equivalence, we can least-squares superimpose the two molecules. (We align by all heavy (non-hydrogen) main-chain atoms.)
3. Then, based on relative positioning of the molecules determined from the fit, we calculate the distance (d) between every atom i in the first structure and every atom j in the second structure.
4. Each distance is transformed into a similarity value to form the similarity matrix shown at **II**:
   $S_{ij} = 20/[1+(d/5)^2]$
   (20 is an empirically-chosen constant; 5, likewise a constant, is for 5 Ang, chosen mostly empirically.) In the initial orientation atom a is close to atom A and even closer to atom C, and this is reflected in the Sij matrix values.



5. Dynamic programming chooses the pairs indicated by the boldface matrix entries. (For the structural alignment portion, no gap penalties are used in the current version of the program; these require significant work to determine their best values, with adjustments to the other parts of the algorithm.)
6. The pairs chosen give a new set of equivalencies (shown at **III**). These are used to do a second least-squares fit.
7. A new similarity matrix can now be calculated, and dynamic programming again used to find new equivalencies.
8. At **IV**, we see that these equivalencies give a perfect match, so a final cycle of dynamic programming does not change the alignment. The iteration has converged on an alignment.

Locally created structural alignments

Locally done alignments were performed using the Structal (Gerstein & Levitt

1996, 1998) method (see Figure 3.2, on page 80), using a locally modified[161] version of the LSQRMS program (Alexandrov & Graham 2003). The matrices used for the initial alignment (prior to the structural portion of the alignment) were as follows:

- Gonnet (Gonnet, Cohen, & Benner 1992)
- Pam120 (Dayhoff, Schwartz, & Orcutt 1978)
- Blosum62 (Henikoff & Henikoff 1992)
- "Nussinov"[162] (Naor *et al.* 1996)
- Identity ("Ident")

All of these matrices were used in an all-positive form, as per the results of an earlier study (Vogt, Etzold, & Argos 1995); the gap penalties were also as per the results of that study[163], as was the choice of the first 3 of the matrices used. The Nussinov matrix was chosen as being derived from a different source (structural equivalence, as opposed to (putative) evolutionary substitutions)[164]; the Identity (Ident) matrix was chosen as a means of lessening bias.[165] In general, these

[161] See patchfile "lsqrms-2.0.4b.patch" for the modifications, or either http://cesario.rutgers.edu/easmith/research/lsqrms-2.0.4b.patched.tar.bz2 or http://cesario.rutgers.edu/easmith/research/lsqrms-2.0.4b.patched.tar.gz for a file containing all program code.

[162] The matrix in question is the M_1 matrix from said paper (the other matrices are not given in the paper; this should not be a problem since none had a significantly higher informational entropy content than the M_1 matrix, the primary one the authors analyzed).

[163] The gap penalties for the Nussinov matrix used those found (Vogt, Etzold, & Argos 1995) for the "pam60_p" matrix, since it had the most similar mean weight, standard deviation, and maximum weight to the all-positive form of the Nussinov matrix.

[164] While the original paper for the Nussinov matrix (Naor *et al.* 1996) indicates that it is not suitable for usage for searches due to its low informational entropy value, this appears to be inapplicable for alignments.

[165] Also tried was a "Glyproalign" matrix of local creation, which was based on the Ident matrix but penalized more substitutions of glycine versus proline versus other amino acids; this was not useful (all examined structural alignments using it ultimately gave the same result as with the Ident matrix, with the only difference being that sometimes more iterations were required). In hindsight, one problem with this matrix was that glycines without unusual phi/psi angles (ones not adoptable by other amino acids) were counted the same as other glycines.

matrices[166] lead to the same result after structural alignment. If not, and it was not obvious that one (or more) matrices had failed (e.g., due to looping between two possibilities) with all the other matrices giving identical results, the best result was judged by a quality measure[167] consisting of the square root of the number of residues aligned, divided by the RMSD. For this measure, higher values were considered preferable.

In some cases, none of the initial sequence alignments were able to determine a rational starting position for the structural alignment (as based on, for instance, very large RMSDs for the starting structural alignment). In such cases, which were unsurprisingly more common the lower the percent identity, some residues were manually determined as being equivalent and their codes altered in the PDB files to force their alignment, and the alignment subsequently rerun (without the Identity matrix and with the "-a" flag to (the locally-modified version of) LSQRMS). Such residues included ones that were:

1. manually determined as being active site residues;
2. other binding site residues (including via examination of ligands in the structures)[168];

[166] The same was true for other variations, namely thresholding of the minimum distance considered "problematic"; please see the source code for more information on this option, which did not appear to make a significant difference, at least in comparison to the choice of matrices, although further analysis would be desirable, particularly in the more difficult cases.

[167] The idea for this measure was derived from the finding that the expected increase in the RMSD is proportional to the square root of the number of residues aligned (McLachlan 1984; Remington & Matthews 1980). If one does not include the number of residues aligned in such a measure, then one can get a (nearly) perfect alignment by simply only "aligning" one residue.

[168] This was the only usage of non-protein (“heteroatom”) locations in the alignments, despite the appearance otherwise of the heme group in some of the hemoglobin alignments, which superimpose almost exactly.

3. conserved tight turns or similar such that only glycine would fit - this was decided by prior data (Lovell *et al.* 2003) on phi/psi angles as compared to those determined via the program “dang” (Word 2000); and
4. conserved cis peptide bonds (generally associated with prolines).

If more than one residue would match one of these categories (most frequently the glycines), then they were either distinguished from one another by the surrounding pattern of secondary structure (e.g., after the first helix - as noted in the PDB files in question - in both structures) and, for glycines, phi/psi angles, or were not used. For instance, the following residues were chosen in ADH1 structure 1CDOA[169] for its alignment to 1HETB[170]):

1. Proline 63, with a cis peptide bond;
2. Glycine 67, with psi 128.37 deg. and phi 106.99 deg.
3. Glycine 86, with psi -10.10 deg. and phi 95.41 deg.
4. Glycine 202, with psi -160.39 deg. and phi -93.52 deg.
5. Cysteine 46, binding the catalytic zinc;
6. Histidine 68, binding the catalytic zinc;
7. Cysteine 175, binding the catalytic zinc.

The corresponding residues in 1HETB were:

1. Proline 62, with a cis peptide bond;
2. Glycine 66, with psi 125.79 deg. and phi 98.47 deg.
3. Glycine 86, with psi -13.53 deg. and phi 95.48 deg.
4. Glycine 201, with psi -164.66 deg. and phi -84.4 deg.

[169] 1CDOA is cod-liver ADH1.
[170] 1HETB is horse liver ADH1, isozyme E (ethanol-active, without activity on steroids).

5. Cysteine 46, binding the catalytic zinc;
6. Histidine 67, binding the catalytic zinc;
7. Cysteine 174, binding the catalytic zinc.

For the PDB files with these alterations, please see http://cesario.rutgers.edu/easmith/research/altered/; these files are distinguished from the normal PDB files by the word "altered" (usually followed by a number corresponding to the set of alterations) included in their names.

Evaluation of structural alignment reliability

Structural alignments were then evaluated for the reliability of different areas of the alignment. Unreliable portions of the alignments were in two categories:

1. Areas of the sequences not found in the PDB files - i.e., intrinsically disordered areas (Le Gall *et al.* 2007) - called "nonstruct".
2. Areas of the sequences for which even structural data were inadequate to determine a reliable alignment, termed "uncertain"[171], decided upon as follows for alignments performed elsewhere:
   - For Pfam, "uncertain" positions were:
     - If the SS_cons or SA_cons line had an "X" or "." for that position;
     - If the seq_cons line had a "." for that position;
     - If the RF line had a gap character (".") for that position;
     - If the sequence differed from that in the PDB file.

[171] Some may make the objection that this is leaving out data. However, failing to take into account structural information about alignments (and other aspects of phylogenetic work) is also leaving out data. This omission is particularly troublesome concerning experimentally determined structural data, since it can potentially provide more information than is currently extractable from the sequences alone (the protein folding problem has not been solved).

- For 3D_Ali, "uncertain" positions had a "-" in the "STRUCT" column, or the sequence differed from that in the PDB file.
- For HOMSTRAD, if the sequence in HOMSTRAD differed from that in the PDB file.[172]

If the alignments were inconsistent between different external alignment sources[173], this was also considered an indication of an "uncertain" alignment if it could not be handled[174] by prioritization (for the above, by 3D_Ali then HOMSTRAD then Pfam). For locally performed alignments, it was concluded that the areas around gaps[175] and at the ends were the most likely locations for uncertainty in the alignment. Therefore, the non-aligned areas of the gaps and ends were "extended" as appeared necessary to give areas of "uncertain" alignments (treated, in essence, as each sequence being versus a gap in the other sequence - thus, the gaps were "extended"). The process for deciding upon these "uncertain" areas was as follows:

1. An initial set of thresholds for maximum distances between superimposed atoms as an indicator for whether overly-distant residues should not be

[172] This criterion is insufficient to detect all truly uncertain locations, but lack of time prevented manual evaluation of each HOMSTRAD alignment, and no other way to determine uncertainty was located for HOMSTRAD.
[173] This included using different PDB files with (essentially) the same sequence.
[174] Such an event would be due to the multiple sequence alignment being from more than one source - see page 93.
[175] One reason for concluding this was the non-usage of gap penalties in the later portions of the structural alignment algorithm. For instance, this resulted in some initial alignments with residues having lengthy gaps both before and after them. These residues were concluded not to be alignable in most cases. Either:
- too many insertions/deletions (Golubchik *et al.* 2007) had taken place in that area, and the structures were no longer truly homologous; or
- there was a region of intrinsic disorder ("nonstruct"), which was problematic if it varied between proteins, whether due to:
  - evolutionary changes (Brown, C J *et al.* 2002); or
  - chance stabilizing interactions between single residues in unstable areas and neighboring

aligned[176] was automatically tried (by the program "ring.changes.lsqrms.pl"), in a process called "ringing the changes" (Adams 1614; Sayers 1934).

2. The results of these thresholds were evaluated by the quality measure explained in footnote 167 (on page 82), as interpreted by the program "interpret.ring.changes.pl".
3. The resulting set of gap "extensions" and associated modified (by the presence of "uncertain" pairings) alignments were manually[177] evaluated and adjusted. Examples of times at which residues would be concluded to be of uncertain alignment included:

---

more-stable areas. Note that it is possible that this could be handled by examining the crystallographic “temperature”/”B-factor”, but this is uncertain (Radivojac *et al.* 2004).

[176] Note that the program was not permitted to remove (on an automatic basis) the pairings of "altered" residues (see under "Locally created structural alignments", on page 82).

[177] It is unfortunate, both in terms of reproducibility and in terms of the time required, that it was not possible to automate this process further. However, it is generally agreed that manual review, at the minimum, is necessary for a good alignment, even (for far-diverged sequences) for structural alignments; hopefully, the further development of artificial intelligence (e.g., computer vision techniques, as used in the creation of the Nussinov matrix (Naor *et al.* 1996)) will enable improvements in this matter. Methods constructing an alignment as part of phylogenetic work, or that essentially examine the possible alignments as part of constructing a tree, are very interesting (Edgar & Sjolander 2003; Holmes & Bruno 2001; Mitchison & Durbin 1995; Mitchison 1999). However, they are so far relatively impractical for large databases and/or not well developed, particularly for protein sequences, in terms of programming (and, as mentioned earlier, fail (thus far) to take into account structural data). The examination of amino acid frequencies from areas concluded not to be reliably aligned, in comparison to the reliably-aligned areas, has given some potentially-interesting deviations (both for "nonstruct" (as might be expected (Coeytaux & Poupon 2005; Penq *et al.* 2005; Penq *et al.* 2006)) and for "uncertain" (Chang, M S S & Benner 2004)), which will be the subject of further work. It is also of interest to note that a survey of the phi/psi angles for the DHFR (3D) structural alignments (conducted using the same methods as for other locally-performed structural alignments), which were not used in the alignments except for tight turn glycines, indicated:

- a close correlation in areas deemed to be reliably aligned; but
- a lack of such correlations for possible alignments in areas deemed to be "uncertain";

again, these will be the subject of further work. Both of these findings indicate that the distinguishing of reliably aligned versus other areas of the alignments was not arbitrary.

a. Positions with large gaps before and after them, due to significant-size insertions and/or deletions, or due to being next to “nonstruct” areas;
b. Positions with significantly differing secondary structures (again, generally due to significant-size insertions/deletions);
c. Positions with residues appearing to have switched places (Azarya-Sprinzak *et al.* 1997), particularly with intermediary stages known (e.g., GX to GG to XG, in which the glycine or glycines are usually needed for a tight turn but their exact position is not strictly limited - the unusual phi/psi angles required could be adopted by residues in either of 2 positions), or with one or the other of a pair of duplicated residues appearing to have been deleted (or, alternatively, a duplicate residue appearing to have been inserted, with no way of knowing which of a pair of residues was the inserted one);
d. Positions which differed in alignment depending on which PDB files (in cases with multiple PDB files with the (approximate) same sequence were available) were used[178]. This situation included if the alignment of sequence A with sequence B, when put together with the alignment of sequence A with sequence C, contradicted (in a multiple alignment) the alignment of sequence B with sequence C.

[178] We attempted to minimize the occurrence of this due to different ligand-binding states by the selection of which PDB files to align, if more than one was available for a given sequence. (For instance, deoxygenated hemoglobin from species A would be aligned to deoxygenated hemoglobin from species B, not oxygenated hemoglobin from species B, even if the latter PDB file would otherwise be preferred by the quality criteria discussed earlier.) It was unfortunately not always possible to do this.

To be noted is that areas considered "uncertain" or "nonstruct" were only considered unaligned with respect to sequences not in the same cluster (of sequences that were 65%+ identical[179]; see item F, on page 27, and “Sequence alignments”, below). Sequence information from these areas was thus not ignored completely for phylogenetic work, only with respect to distant sequences. (In other words, “uncertain” and “nonstruct” areas were aligned to, and used for phylogenetic work with respect to, other sequences that were of sufficient percent identity as to be validly alignable by sequence. They were not aligned to, and not used for direct[180] phylogenetic work with respect to, other sequences that were too far away to be validly alignable by sequence.)

Sequence alignments

Following the construction of the structural alignments, sequence alignments to the structurally aligned sequences were performed; these used only sequences that, according to `blastp` (see "Structures and sequences", on page 61) were at least 65% identical to the structural sequences - in a "cluster" around the structurally aligned sequences. For an example of clusters, please see the table on page 89, remembering that each sequence in a cluster is at least 65% identical to the sequence of at least one (3D) structure in that cluster.

[179] Sequences that are 65%+ identical should be alignable by sequence alignment with equivalent results to structural alignment (Vogt, Etzold, & Argos 1995). Thus, unless one was to put in “uncertain” areas for sequence alignments, one should not make structural alignments *less* certain than sequence alignments when sequence alignments would be valid.

[180] In this, “direct” means as part of the sequences input into MrBayes, not as used to determine, for instance, state frequencies (see “Partitions: State frequencies”, on page 107).

Protein	Cluster	3D Structure(s)	Sequence	Species
ADH1	1HETB	1HETB 8ADH0	**ADHE_HORSE**[181]	**Horse**
		1HSOA	**ADH1A_HUMAN**	**Human**
		1U3UA	**ADH1B_HUMAN**	
		NONE	**ADH1_STRCA**	**Ostrich**
			XP_535667	**Wolf/Dog**
			ADH3_COTJA	**Quail**
	1CDOA	1CDOA	**ADH_GADCA**	**Cod**
		NONE	**Q6B4J3_ORYLA**	**Medaka**
			Q90Y38_BRARE	**Zebrafish**

These alignments were initially by the program "align.to.central.2.pl"[182] using "needle" from EMBOSS (Bleasby 2000; Rice, P, Longden, & Bleasby 2000), using all 5 of the above matrices. If at least 3 of the 5 matrices gave the same result, then this was used. If not, the quality of the alignments was evaluated using several criteria[183]:

1. What the percent identity was of the aligned residues;
2. What proportion (of the maximum possible) of the residues were aligned;
3. What matrix (or matrices) had worked the best with that "cluster" (group of 65%+ similar proteins);

If the above were uncertain, and the alignments gave less than 65% identities, then the sequence to be aligned was excluded. If the above were uncertain but some of the alignments (other than Identity) gave more than 65% identities, then the program "combine.align.to.central.3.pl" was used to attempt to create a

[181] Note that ADHE_HORSE is now identified in SWISS-PROT as ADH1E_HORSE.
[182] Other programs were also used, some of which are mentioned below; also see the supplemental file "Makefile.prior.txt" (also available via http://cesario.rutgers.edu/easmith/research/perl/Makefile.prior.txt)- the programs necessary are also available as supplemental files and via a webserver (see “Appendix P: Perl programs created”, on page 415).
[183] These criteria were based on evaluations of alignments with the matrices versus structural alignments (of two structures at a time), using only alignments of structures with 65%+ identical sequences (these did not, or at least should not have, *required* structural alignments to be usable;

compromise alignment, taking what areas were agreed upon by several matrices[184]. If this was not possible, then the sequence to be aligned was excluded.

Alignments were also performed between "nonstruct" areas of otherwise (3D) structurally known sequences and other sequences in the same cluster. This used matrices, as with the above, plus Pfam (since it is actually a sequence alignment database, strictly speaking). The program to perform this was "check.pdb.vs.pfam.pl".

Multiple alignments: Inter-cluster

The sequences in the clusters, aligned versus one or more (3D) structurally known sequences in the cluster, were then put together into a multiple alignment. What does one do when one has gaps in the (3D) structurally known sequence versus two or more non-structurally-known sequences? Please see Figure 3.3, on page 91, for the “xgap” algorithm used.

---

the alignments were done in order to evaluate the criteria). See "Appendix H: Evaluation of alignment quality", on page 375, for the programs involved in these evaluations.

[184] Matrices that appeared to have reasonable results by the above criteria were emphasized, as were areas that had agreements between 3 or more matrices; the algorithm used was to start with looking at the results from all matrices for agreements, then take out matrices in order of estimated quality to resolve the remaining areas that were in dispute between matrices.

```
Structural sequence: AAA---CCC
Other sequence 1:    AAApgeCCC

Structural sequence: AAA--CCC
Other sequence 2:    AAApdCCC

Structural sequence: AAA---CCC
Other sequence 1:    AAApgeCCC
Other sequence 2:    AAAp-dCCC
```

The non-structurally-known sequences in the above are aligned to each other via a further sequence alignment - on top of the alignment that has already been done to the (3D) structurally known sequence. (The algorithm involves preserving existing gaps via changing them (temporarily) into "x" characters - thus the name "xgap".)

Figure 3.3: Xgap algorithm

However, the usage of the "xgap" algorithm was avoided when possible, such as by aligning versus a structurally known sequence that did not have a gap in the area in question. One problem with the "xgap" algorithm is the question of what order to align sequences in - if there were more than 2 non-structurally-known sequences in the above, then there would be a dilemma about which to align the others to, for instance. We used, for each gap, the sequence with the most characters in the gap first (with ties between the number of characters broken by the quality of the original alignment), so that the other sequences would have residues to align against. Gaps were only added where necessary - other sequences were tried first, if possible, if a result indicated a new gap would be created. Another standard used for alignment quality in this was whether the alignments in question turned out the same if the sequences were reversed[185]. Involved in this process was that sometimes, more than one structurally aligned sequence was 65%+ identical to a given sequence without a known (3D)

[185] This idea originated prior to seeing an interesting recent article suggesting the usage of this technique in evaluating the reliability of alignments (Landan & Graur 2007).

structure. The alignments could differ between these. Which is given priority? For sequences individually, we prefer:

1. Alignments in which multiple matrices gave the same result;
2. Alignments not requiring the use of the “xgap” algorithm.

If the above did not work, for each cluster, the possibilities were sorted by the following criteria, in order:

1. Had been locally structurally aligned (as per "Locally created structural alignments", on page 80) to structures outside of its cluster;
2. Had been aligned by more than one method (e.g., both 3D_Ali and HOMSTRAD) to structures outside of its cluster;
3. Had been structurally aligned (by any method) to structures outside of its cluster;
4. Had the most proteins in its cluster;
5. Had the most proteins alignable to it;
6. Had the best structure in its cluster (by the criteria used by “interpret.important.pdbs.pl” - see “Appendix B: Important PDB files/chains used”, on page 367);
7. Had an identical sequence to the one in the databank for its origin species;
8. As a fallback (which was not, as far as we are aware, necessary), alphabetical order.

The above was performed by "integrate.sequence.align.1.pl".

Multiple alignments: Structural

Following the above, the structural alignments, previously in pairwise alignments, were put together into multiple alignments by the program "integrate.structural.align.1.pl". If one has an alignment of X versus Y, and one of Y versus Z, it should be possible - and is desirable - to generate an alignment of X versus Z (which can be considered to be from *both* the method used for X versus Y (e.g., Pfam) *and* the method used for Y versus Z (e.g., 3D_Ali). However, one could also have an alignment of X versus A and A versus Z, and the resulting alignment (of X versus Z) from it could conflict. For this situation, the alignments were prioritized by method:

1. Locally-performed structural alignments;
2. Matrix alignments, for 65%+ identical sequences only (i.e., when two structures happened to be in the same cluster);
3. 3D_Ali;
4. HOMSTRAD;
5. Pfam.

When (usually due to derivation from multiple sources) this was uncertain, the program created a compromise (Lake 1991) alignment (in which, if applicable, conflicting areas were considered not reliably structurally aligned, denoted as "uncertain").

Further sequence processing: Ambiguity-coded polymorphism reduction

Due to the presence of significant amounts of ambiguity coding in the compromise sequences created by the procedure in “Usage of polymorphism”, above, combined with MrBayes’ limits in handling ambiguity-coded polymorphism[186], it was concluded that it would be desirable if the amount of ambiguity coding were decreased[187]. (This decrease would be after information had been gathered on amino acid frequencies; see "Partitions: State frequencies", on page 107.) The intended result of this was for the compromise sequences found in the lowest numbered[188] "full" species to be closer to the sequences found in either:

1. all other species (if possible); or
2. other species that appeared likely[189] to be phylogenetically close to the species in question.

This reduction took place via the program "nexus.simplify.polymorphism.pl". In this program, the species groups were gone through, from the smallest group encompassing the species to the largest group, containing all species - e.g., "plant/algae" then "eukaryota" then "all" - for *Arabidopsis thaliana* - with the amino acids[190] present in the other species (in the group) being checked for

[186] However, some work has gone into improving this - see item 2 on page 98.
[187] This was a further decrease after that in “Species, polymorphism reduction”, on page 70, enabled by the alignment’s inclusion of other species’ sequences.
[188] Please see ‘Creation of “full” species’, on page 68.
[189] This was based on phylogenetic groupings that were not considered to be in dispute, such as bacteria versus archaea versus eukaryota. For other examples, see "Appendix I: Species groupings used", on page 376; only clade groups were used for this.
[190] Instead of amino acids, this could be the gap coding "DNA" - see "Gap determination", on page 139.

identity or similarity[191] to those in the ambiguity coding. The first "smallest" group for which a given identity/similarity definition gave an intersection (overlap) with that from the ambiguity coding[192] was used.

The possible identity/similarity definitions can give rise to different intersection groups; if so, the result was chosen (to replace the original ambiguity coding) that was best by the following criteria, in priority order:

1. Giving only one possibility, thus eliminating ambiguity coding;
2. Being identical (not simply similar) to the original, with using all species instead of only the first of "full" species (see "Creation of “full” species", on page 68);
3. Having the lowest number of possibilities for the ambiguity coding;
4. Using the strictest criteria for similarity/identity to the original;
5. Using all species instead of only the first of the "full" species.

To be noted is that this process actually ran in two stages; the first stage used the smaller groups, while the second stage also used the "all" group (with all species with the protein in question). This method was partially for group sequence creation (see below) and partially so that the second run could take advantage of narrowing by the first run.

[191] Similarity was done by the definitions found in "Appendix G: ESIMILARITY matrix", on page 374, for amino acids. For the gap-coding "DNA", an initial guess was modified in light of the "sump" results (see "Usage of the results of prior tree runs", on page 127) for the GTR (see "Gap determination", on page 139) for the gap-coding scheme - please see the program code for the initial matrix and the one used later.

[192] For instance, if the other species had "A" or "C", and the ambiguity coding had "C" or "T" as possible, then "C" would be the intersection; this is the intersection (overlap) between two sets (Wikipedia 2007).

Further sequence processing: Group sequence creation

It was necessary to reduce some groups of species' sequences[193] to single sequences for reasons of computational load. The groups to which this applied included both:

- outgroups (see "Appendix O: Outgroup review/explanation", on page 412) and
- some internal (to the group of species of interest) but less important (and/or inadequately represented for accuracy) groups of species (see "Appendix I: Species groupings used", on page 376, for the groups used)

Initially, group sequences were created by "nexus.use.recdcm3.subsets.pl", and were simply created by putting together all residues found at a given location for any species in the group as the "polymorphic" alternatives. This process resulted in a considerable degree of ambiguity. After:

- the difficulties with MrBayes and ambiguity were realized (see "Further sequence processing: Ambiguity-coded polymorphism reduction", on page 94); and
- better trees with reasonable distances were derived,

it was decided to create group sequences using weighting[194] of sequences, together with using not only identity, but also similarity (see footnote 191, on page 95, for more information). This process was performed by the program "create.outgroup.seqs.pl", using "full" species (see "Creation of "full" species", on page 68) narrowed down to single "real" species sequences (with ambiguity

[193] These are referred to in some material as "outgroup sequences" from "outgroups", since outgroups like (for Eukaryota) Bacteria or Archaea were most frequently used to create them.
[194] The weighting was as per "Alignment using HMM", on page 129, but using the program "find.species.weights.2.pl", since the intended weightings were only of the outgroup sequences.

removed when possible - see "Further sequence processing: Ambiguity-coded polymorphism reduction", on page 94), by the programs "nexus.simplify.full.species.pl" and "nexus.simplify.full.species.2.pl". The groups used were also a more limited set of groups (see "Appendix I: Species groupings used", on page 376).

The above process narrowed down the ambiguity at each sequence location to no more than two possibilities. A single possibility was used when possible (when one residue or equivalent made up over 50% of the weight, for instance).

Species considered in the group for later processing[195] were solely those with sequences contributing amino acids (or equivalent) to the group sequence in question. These were narrowed down further by "nexus.use.recdcm3.subsets.pl" via its elimination of some positions (for instance, positions with only one species in a subset having a residue present were removed from that subset's sequences), although perhaps not as much as would be possible (due to time constraints).

[195] This would be primarily for the determination of tree distances - see "Tree distances", on page 113. Note that more than one sequence (e.g., from more than one protein) could come from a given species, and the determination of branch lengths uses all of the sequences, with some proportionality (rate variation; see “Partitions: Gamma, Invariant, Rate”, on page 105, and under “Tree distances”, on page 126). It is thus important to keep track of, not simply the sequence resulting from the group combination, but the species contributing to them.

4. Tree refinement

MrBayes code alterations

Tree refinement was carried out primarily using the program MrBayes (Huelsenbeck & Ronquist 2001; Huelsenbeck *et al.* 2006; Ronquist & Huelsenbeck 2003; Ronquist 2005). The MPI (parallel processing) implementation of MrBayes (Altekar *et al.* 2004) was tried, but was found to have reliability[196] and error diagnosis problems. After this was determined, parallel runs of MrBayes were initialized by hand on different machines/processors, not automatically, and comparisons between runs were initiated manually[197]. Some modifications to MrBayes were made (see patchfile "mrbayes-3.1.2.patch"); these were in several categories:

1. Those needed for compilation (including optimization, error-checking, and debugging) on IRIX[198] and Linux;
2. Those affecting the assumed state (e.g., amino acid) probabilities when polymorphism or other uncertainty (missing data or gaps) was present. The original code in MrBayes set the probabilities of each of the possible states

[196] The reliability issues were probably related to inter-program communication problems, which can unfortunately be extremely difficult to sort out with parallel programming.

[197] MrBayes also has the capability of running, particularly in parallel, more than one "chain", with "swapping" between these chains. In this process ("Metropolis Coupled MCMC"), some chains will be run with less strict criteria for the acceptance of "moves" than other chains, so that they can explore more possibilities, with some data communicated between chains ("swapping") when they were sufficiently successful (Altekar *et al.* 2004). (See Appendix J: MrBayes review/explanation", on page 379, for more information.) However, it was found for our dataset that either very little chain swapping took place or the chain differences in "temperature" (ability to explore more possibilities; see item 6 on page 100) were very small (e.g., see "Tree search with Eukaryota (subset)", on page 300). Therefore, multiple chains were not used for later runs and simulated annealing (again, see item 6 on page 100) was used instead for the purposes of chain swapping and temperature increases.

[198] IRIX is the SGI (Silicon Graphics) variety of UNIX.

(e.g., amino acids) as being equal (to 1 each) even if there was more than one state possible. This coding appeared problematic for two reasons:

a. It meant that the total probabilities would add up to more than 1 (at least prior to any scaling) if there were multiple residues possible for a position in a sequence;

b. It failed to take into account that some residues are more likely than are others (e.g., tryptophan is less common than alanine).

Accordingly, the (local version of the) code was altered so that the probabilities at positions with multiple residues (or nucleotides, *etc.*) instead added up to 1 with these being distributed among the possibilities in proportion to the state frequencies[199].

3. Those involved in attempting to get the "covarion"[200] option to work, including diagnosis of various problems encountered;

---

[199] In cases in which the state frequency estimates (see "Partitions: State frequencies", on page 107) are subject to alteration ("moves"), namely with Dirichlet state frequencies, it would admittedly be advisable to redistribute the probabilities whenever such a "move" took place, but this was found to be too complex to implement in the time available. Another difficulty is that some residues are more associated with gaps than others are, as noted previously (Chang, M S S & Benner 2004); it would be preferable to adjust the proportions assumed for gaps in consideration of this, but this was again found to be too complex to implement in the available time. For testing and for usage by others, it will also be preferable to make whether this modification is used switchable (ideally on a partition-by-partition basis). For further notes on testing of this, please see “Discussion and future work”, on page 344.

[200] In the covarion option in MrBayes, positions are assumed to vary (along the tree) between being variable and invariant. Unfortunately, the usage of the covarion option with our dataset resulted in significant errors (particularly if used with gamma rate variation), including:

- LIKE_EPSILON (numbers too close to zero, indicating probable roundoff errors) error messages - see footnote 423 under “4. Tree refinement”, on page 195, for more information;
- extremely low proportions of "moves" (see “Appendix J: MrBayes review/explanation”, on page 379) altering covarion proportions (likelihoods of positions going from variable to invariant or vice-versa) being accepted; and
- probabilities going significantly above 1 (positive log probabilities).

It appears likely that the covarion option works better with (and was probably primarily or entirely tested with) nucleotide data, given that it effectively expands the transition matrix for amino acids to 40x40 (from 20x20) but the nucleotide matrix would "only" go from 4x4 to 8x8. Note also that this is actually a "covarion-like" model, in that the original covarion model has some level of dependence between sites in whether they are variant or invariant, whereas the model

4. Those involved in various internal improvements to its functioning, such as with regard to "scaling" to avoid roundoff errors (the change made also increased the speed of the program, since scaling was only done when it was detected that roundoff errors would otherwise occur);
5. Those involved in changing the "props", or proportions of "moves"[201] tried, including a "notopology" mode (enabled by compilation with a "-DNOTOPOLOGY" flag) in which (as per the MrBayes user manual's suggestion) no moves were made that changed the topology (branching pattern, as opposed to, for instance, distances) of the tree;
6. Those involved in putting in Simulated Annealing (Kirkpatrick, Gelatt, & Vecchi 1983), abbreviated "SA", as a replacement for MrBayes' Metropolis Coupling of Monte Carlo Markov Chains (Altekar *et al.* 2004). In this technique, the "temperature" is at first high (resulting in a high probability of acceptance of "moves", to avoid being trapped in a local minimum). The temperature[202] is then lowered - to zero *prior* to the end of the (minimum expected) "burnin"[203] phase, after which samples are to be gathered - to try to locate the most likely possibility (or set of possibilities). Please see “Adapt and SA”, on page 381, for more information.

---

implemented in MrBayes has sites independently switching between states (Galtier 2001; Huelsenbeck 2002; Huelsenbeck *et al.* 2006; Tuffley & Steel 1998).

[201] Please see “Appendix J: MrBayes review/explanation”, on page 379, for more on “moves” and MrBayes.

[202] Note that the "temperature" that MrBayes shows is the inverse of this - a chain that is not "heated" in MrBayes will be shown as having a "temperature" of 1.0, while chains that are more free to vary will have a lower "temperature". Again, for further on “temperature”, please see “Appendix J: MrBayes review/explanation”, on page 379.

[203] Please see footnote 428 under “Simulated Annealing (SA)”, .on page 197, and “Adapt, SA, and burnin”, on page 383.

7. Those involved in adding adaptation (Corana *et al.* 1987) of "move" sliding window/multiplier sizes (abbreviated as "Adapt"), as a partial replacement for the rather difficult process of tuning said parameters for a given dataset via "props"[204]. (It is recommended (Huelsenbeck *et al.* 2006; Ronquist 2005) that the acceptance rates of the "moves" be between 10-70% (ideally, 20-70%) for optimal usage, but this can require a considerable degree of tuning after seeing the results of program "runs" with a particular dataset.) As with the simulated annealing code, the adaptation will halt prior to the end of the (minimum expected) "burnin" phase, to avoid disturbing the "run" during the data-gathering portion. (Again, please see “Adapt and SA”, on page 381, for more information.) This portion of the code changes included those resulting from the finding that some of the "reflection" code in the moves (which reverses the direction of a move if it goes outside the allowable range) was capable of entering an infinite loop with some window values. Code to detect this (by that an already-reflected move appeared to need to be reflected again) and abort the move was inserted.

Species subsets

As previously noted (see under “Need for starting tree”, on page 31), the number of species in use necessitated not using the entire set of species at once[205].

[204] Note that the “props” command is considered an advanced area in MrBayes 3.1.2 (Huelsenbeck *et al.* 2006), and indeed the setting of “props” for automated runs necessitated alterations to the source code, at least in this version of MrBayes (Huelsenbeck *et al.* 2007).

[205] The exception was for the initial determination (without DHFR sequences) of approximate branch lengths, with the tree topology remaining fixed. This process required the usage of a

REC-I-DCM3 (Huson, Nettles, & Warnow 1999; Roshan *et al.* 2004a; Roshan *et al.* 2004b) was chosen as the method to break up the tree into subsets. This method (Recursive Iterative Disc-Covering-Method-3) breaks a tree up into overlapping subsets that together span the entire tree but that individually minimize the distances inside each subset (generally with an overlapping set of divider species that are close to the center of the tree). This method, as well as making it more possible to handle large species sets, can improve accuracy[206]. Three difficulties were found with REC-I-DCM3 in its downloaded (1.0) version[207]:

1. As mentioned previously (see under “Need for starting tree”, on page 31), REC-I-DCM3 requires a fully resolved tree. While it is apparently capable of converging on a reasonable tree even if started with a tree that is not very accurate (e.g., one for which polytomies[208] were “resolved” by being randomly split up), the time required for the repeated tree searches needed to resolve such would be significant on a dataset the size of ours. Thus, it was desirable to create a fully resolved tree to start with.
2. As mentioned earlier, not every protein sequence is known (or is even present) for any single species; some groupings of species would not be phylogenetically useful due to a lack of known protein sequences in common. REC-I-DCM3's generated subsets are based entirely on the tree

---

computer with well over 2 GB of memory and a considerable amount of time.

[206] The improvement in accuracy is because many phylogenetic methods are less accurate over very long evolutionary distances. Among the reasons for this inaccuracy are overlapping mutations, which are difficult to distinguish from single mutations giving rise to the same result.

[207] Updates to the program since then, which unfortunately appear to have taken place mostly or entirely after this portion of the research was complete, may have fixed one or more of these problems. The program authors will be notified of any problems that remain.

and its distances. These subsets thus may[209] include such groupings, or at least groupings in which inadequate proteins are known for reliable phylogenetics.

3. REC-I-DCM3's subset generation is focused on making sure that the subsets do not include species that are too distant from one another, insofar as this is possible. However, for purposes of generating the smallest useful subsets, it is also helpful to consider whether species are too close to each other (e.g., polymorphism-generated "species" as noted under “Usage of polymorphism” on page 64, or such pairings as *Homo sapiens* and *Pan troglodytes*).

With regard to problems 2 and 3, the program "recdcm3.get.subsets.pl"[210] was constructed to generate subsets of species for further phylogenetic work. This program has two modes:

1. The one used for the initial few rounds of tree refinement; in this, REC-I-DCM3 is run to generate subsets of various sizes (including subsets further split up than is desirable for actual usage). These are then combined into or split between subsets better fitting the criteria;
2. The one used for later stages; in this, REC-I-DCM3 is not actually used, with the initial subset(s) coming from an (internal to the program) input of

---

[208] Polytomies are places in which a (non-root) node has more than 2 descendant branches.

[209] One factor causing them to be less disconnected is that the distribution of some proteins (e.g., myoglobin, found solely in metazoa) is correlated with the phylogeny. Unfortunately, this is not sufficient to ensure usability of all (or even most) subsets - particularly for a multi-(super)kingdom set of species like that of the current study - for a variety of reasons. These reasons include that, even for proteins whose presence is correlated with the phylogeny, the *sequence* of said proteins may not be known (or may not be close enough to be alignable even if known).

[210] It is possible that the program in question contains one new algorithm in graph theory, namely a relatively low-complexity (approximately linear in the number of nodes involved) method to do

species of interest (e.g., fungi/metazoa with DHFR, mammals, or fungi), and the generation of more subsets from these subsets based on distances to other species.

Of importance in regard to subsets of species with usable proteins are the prior findings of the minimum number of species (sequences) necessary for adequate determination of the alpha parameter for gamma (rate variation; see footnote 215, on page 105) for each protein. This finding was of 20 at a minimum, and 30 if more than 4 rate categories are deemed necessary (Blouin, Butt, & Roger 2005; Meyer & von Haeseler 2003; Pollock & Bruno 2000). When possible, the program uses subsets with as many proteins as possible meeting these criteria[211]. Further manual[212] and automatic[213] modifications to the subsets generated by this program were necessary at times. (Some good examples of

---

the equivalent of REC-I-DCM3's splitting for a flat graph instead of a tree; further checking on this is desirable.

[211] Moreover, for a protein to be counted as usable at all for a subset, at least 4 species in the subset must have known (alignable) sequences for it, since less than that number will not determine even a quartet.

[212] Manual modifications were primarily to do one or more of:

- reducing the number of species to a more manageable number, generally along with one of the below;
- combining two (or more) subsets that were largely but not entirely equivalent; and
- adding polymorphic DHFR sequences for ancestral sequence determination runs.

[213] Automatic modifications, by the program "nexus.use.recdcm3.subsets.pl", were partially based on manually-added rules (taking into consideration whether species would overlap with others in terms of sequences known) from:

1. prior research (Anderson & Swofford 2004; Gibb *et al.* 2007; Graham, Olmstead, & Barrett 2002; Moreira, Lopez-Garcia, & Vickerman 2004) on breaking up long branch lengths with the addition of species;
2. problems encountered with distance determination (see "Tree distances", on page 113);
3. problems (inconsistent placement) seen with earlier tree searches (as found by "compare.trees.problems.pl").

The other automatic modifications were from the substitution of group sequences (see "Further sequence processing: Group sequence creation", on page 96) for species if very few species in said outgroups were present;. This substitution was done somewhat more (as in, more stringent criteria for *not* substituting an outgroup sequence, particularly for species without DHFR or DHFR/TS sequences usable/known) in later rounds, due to greater confidence in the improved outgroups and a desire to focus on, for instance, Eukaryota (since the DHFRs used are from eukaryota) at later stages.

subsets can be seen in the trees under “Second round of tree rearrangements”, on page 265, and “Tree searches”, on page 299.)

Partitions: Gamma, Invariant, Rate

It was important to consider the number of species/sequences per protein (or section of protein - see below) not only for subset determination, but also for the creation of the input files for MrBayes. Initially, the proteins were divided into "partitions"[214] based on their categories of alignment - namely structurally aligned ("struct"), "nonstruct" (intrinsically disordered (Le Gall *et al.* 2007)), and "uncertain" - and, for those categories not structurally aligned ("nonstruct" and "uncertain"), different clusters (65%+ identical, outside of which only structurally aligned residues were considered aligned). In some cases, these partitions had one or more problems:

- With too few species;
- With too short sequences; or
- With too little variability.

These problems were of concern for the reliable determination of:

- the alpha parameter of the gamma rate distribution[215];

[214] The term "partition" is used in MrBayes to indicate a set of sequence or other data that has been divided from other sets of sequence/whatever data so that it can be treated differently (not as part of the same sequence).

[215] The "alpha" is a parameter that adjusts the model for variations in the rate of evolution (e.g., sequence changes) for different positions in a sequence. It alters the shape of a curve following a "gamma distribution" (please see http://en.wikipedia.org/wiki/Gamma_distribution for an example picture) to indicate the probability of a site (position) having a given rate of evolution. In MrBayes, this is done - as with most phylogenetic software - by dividing sites into several discrete rate categories. For instance, one might have a category with 0.25 mutations per evolutionary distance “unit” (an average of one mutation per location, in general), a category with 0.5 mutations per distance unit, a category with 1 mutation per distance unit, and a category with 2 mutations per distance unit. Each site would have a probability of association with each of these

- the invariant proportion; and
- the relative rates[216] between partitions.

One way to look at this question is the avoidance of overparameterization - too many parameters for the available data to properly fit (Huelsenbeck *et al.* 2006; Kjer 2007). The minima (minimums) used were:

1. for gamma, 20/30 species/sequences as noted on page 104;
2. for invariant proportion[217], determined by the amount of data needed for a binomial 95% confidence interval of +/- 0.5 or less;
3. for rate determination, a guess primarily based on the minima for gamma and invariant proportion

When necessary, the partitions for a given protein were merged ("linked"[218]) for one or more of these parameters (with no merger with "struct" if both "nonstruct" and "uncertain" were present), if they would otherwise be likely to have

---

categories in terms of its frequency of (evolutionarily accepted) mutations. How different the categories are from each other is determined by “alpha”; in the version of MrBayes used (3.1.2), each site has an equal probability of being in each category. Note that these rates (from the gamma distribution) are relative to each other within a specific partition, not relative to other partitions (see under “Tree distances”, on page 126). Note also that in the present work, 4 categories were used unless there was clearly enough data (e.g., enough sequences - 30 or more - and enough variation in these sequences) to use more (Blouin, Butt, & Roger 2005; Meyer & von Haeseler 2003; Pollock & Bruno 2000).

[216] See under “Tree distances”, on page 126.

[217] The invariant proportion (“pinvar”) is the proportion of locations in the sequence that are invariant (constant/fixed). In MrBayes, it can be determined either along with the “alpha” for “gamma” in the “invgamma” model, or without “gamma” in the “propinv” model. (The “covarion” model is a replacement for it allowing for variation along the tree - however, see footnote 200 under “MrBayes code alterations”, on page 99.) In the present work, the maximum proportion invariant was set to 1 minus the proportion of variability seen for the partition in question, with the minimum being 0 (since positions that currently appear invariant could actually have been variable in an ancestral sequence).

[218] Parameters that are “linked” between different partitions are constrained to remain the same, thus using data from both partitions to determine what they should be. It should be noted that MrBayes (version 3.1.2) does not have the capability of linking rates between different partitions. Ultimately, such cases had to be handled by the complete merger of the partitions in question (by "nexus.consolidate.partitions.pl", run on the products of “nexus.use.recdcm3.subsets.pl”). Given that this merger was avoided when possible as eliminating data, this limit of MrBayes may have been partially responsible for the problems seen with between-partition rate variations; see under

significant problems. (In a few cases, there was only one sequence in a given partition[219], in which event the partition was deleted.) Please see the program "combine.structural.align.groups.pl" for more information on this stage, including exact details on how the minima were created and used; the supplemental file "Makefile.txt"[220] contains information on the origins and usage of the files and other programs (keeping in mind that DHFR sequences were added manually - see "Usage of polymorphism" on page 64).

Partitions: State frequencies

One parameter of phylogenetic models is what the background frequencies[221] are of the "states" (in our case, amino acids) in the dataset in question, including whether to keep this static or allow it to vary as one of the parameters (Dirichlet state frequencies). This parameter may appear simple ("just" use the existing (visible) frequencies). However, as well as having to take into consideration polymorphism (see "Usage of polymorphism", on page 64), a distinct problem can be encountered if we lack sufficient sequence data. In its most extreme form,

---

"Tree distances", on page 126.

[219] I.e., there was a protein with only one species in a cluster (that was used) and that protein had a "nonstruct" or "uncertain" section, not usable outside the cluster

[220] This file is used by the program `make` (Stallman, McGrath, & Smith 1998) to direct the second part of the sequence processing.

[221] The “background frequencies” are the proportions (frequencies) to be assumed for purposes of phylogenetic modeling of the “states”, such as amino acids. This influences, for instance, what the likely ancestral amino acids were of a given amino acid (less common residues are, after allowing for substitution matrix differences, less likely to be what a residue mutated from due to being less likely to have happened in the first place). One matter making the automated variation of this parameter (or set of related parameters, depending on how one looks at it) more complicated is that the total of the proportions ultimately used must add up to 1 - a decrease in any one amino acid’s frequency implies an increase in another amino acid’s frequency, for instance. It should be noted, incidentally, that using MrBayes’ built-in matrices results in a matrix that is not adjusted for state frequencies; this was overcome in the present research by entering the WAG (Whelan & Goldman 2001) matrix as a fixed GTR (for which state frequencies are properly adjusted).

if the extant examples of a protein[222] entirely lacked an amino acid (e.g., tryptophan), to use the existing state frequencies would imply that the protein not only currently lacked that amino acid, but also had also *always* lacked it. In other words, it would imply that there is no chance that this amino acid was present in another sequence (of that protein) and, more importantly for the present work, no chance this amino acid had been present in the past (Durbin *et al.* 1998). Another difficulty is that MrBayes' "move" (see Appendix J: MrBayes review/explanation", on page 379) with regard to state frequencies, Move_Statefreqs, will not accept anything lower than 0.01%[223] for use with Dirichlet proportions (which means that a somewhat higher level is advisable to allow for, for instance, roundoff error).

For dealing with partitions that were overly small, the possibilities were as follows:

- Combining one cluster's "uncertain" partition with its "nonstruct" partition;
- Combining one cluster's "uncertain" partition with another cluster's "uncertain" partition;
- Combining one cluster's "nonstruct" partition with another cluster's "nonstruct" partition;
- Combining an "uncertain" or "nonstruct" partition with the protein's "struct" partition - this was avoided when possible, given both the differences

[222] This consideration would also be for a section (one intended to be used as a partition) of a protein, such as the "uncertain" or "nonstruct" partitions for a particular cluster. The example in question (the entire lack of an amino acid) did happen in some cases of this, due to the short length of the partition and/or the low number of species.

[223] Altering this number was contemplated, but decided against on the grounds of avoidance of

observed in the present research between these areas and the “struct” areas and prior research likewise indicating the presence of such differences (Chang, M S S & Benner 2004; Coeytaux & Poupon 2005).

The above possibilities were decided between, by the program "nexus.add.freqs.pl", via a combination of criteria including[224] informational entropy, (an estimate of) the Bayes error[225], and informational loss (Lin 1991; Yona & Levitt 2002).

There were several choices available for dealing with the problem of overly low frequencies (including of the newly combined partitions from the above):

- Combining frequencies[226] from one cluster's "uncertain" partition with that of its "nonstruct" partition;
- Combining frequencies from one cluster's "uncertain" partition with that of another cluster's "uncertain" partition;
- Combining frequencies from one cluster's "nonstruct" partition with that of another cluster's "nonstruct" partition;
- Combining frequencies from an "uncertain" or "nonstruct" partition with the protein's "struct" partition's frequencies - this was avoided if possible;

---

cumulative roundoff errors potentially causing values to go too close to zero.

[224] Such a measure as chi-square would not be suitable for this, since what it is measuring is not how important the deviations are between two sets, but how likely it is that the differences seen are due to chance; these are two different questions.

[225] Unfortunately, some research (Wu, T D, Nevill-Manning, & Brutlag 1999) indicating the squared difference between proportions ("squared error") to be of use was not realized to be using a different error measure than Bayes error, and thus of importance to read as not duplicating earlier information, until too late to use this method.

[226] By "combining frequencies" between partitions, putting together the frequencies and linking the state frequencies for the partitions (see footnote 218, on page 106) is meant.

- Combining frequencies with those used in HMMer (Eddy & Birney 2003) for "insert" state proportions (these are for areas that are considered unalignable with HMMer) - again, this was avoided if possible.

The above possibilities were decided between, by the program "nexus.add.freqs.pl", via a combination of criteria including (an estimate of) the Bayes error and the Kullback–Leibler divergence (Lin 1991; Liu, X Z *et al.* 2003; Yona & Levitt 2002).

Another question concerning state frequencies is whether to use them as a fixed quantity ("fixed statefreqs") or whether to allow them to vary via MrBayes "moves" ("Dirichlet statefreqs"). This question was decided upon, by the program "nexus.add.freqs.pl", by a variety of criteria:

- If, despite the above, one or more of the statefreqs were too low for MrBayes' Move_Statefreqs to handle, then fixed statefreqs were necessary;
- If the statefreqs (or the partitions) had overly-low statefreqs, but not so low that Move_Statefreqs could not handle them, then Dirichlet statefreqs were preferable;
- If the partition (or group of partitions) was, despite the above, smaller than desirable, then fixed statefreqs were preferable;
- If the statefreqs (or the partitions) had been grouped by the above, then Dirichlet statefreqs were preferable;

- If the statefreqs were from sequences that had earlier been removed (see "Species, polymorphism reduction", on page 70), then fixed statefreqs were preferable;
- If there were more than 300 amino acids[227] in the partition in question, fixed statefreqs were preferable;
- If the number of amino acids was less than the number of species involved, then Dirichlet statefreqs were preferable;
- If none of the above criteria were true, then state frequencies were fixed, in order to reduce the number of parameters involved (i.e., avoid overparameterization; see "Partitions: Gamma, Invariant, Rate", on page 105).

Tree rearrangements

It was unfortunately found that, even with only subsets of species in use and a starting tree[228], doing a full tree search using MrBayes was not practical for most subsets tried[229] due to time constraints; different program "runs" did not converge on the same tree in any reasonable amount of time. This problem appeared to be primarily because the initial perturbations, and much of the later attempts at rearrangement of the tree by MrBayes (likewise on a random basis), were not particularly likely to be valid[230], and thus for MrBayes to happen upon something

[227] In the MrBayes source code (Huelsenbeck *et al.* 2006), by default the importance of the initially input Dirichlet state frequencies is set to be equivalent to 300 amino acids.
[228] Perturbations (random rearrangements of parts of the tree) were used to get alternate possibilities.
[229] See "Tree searches", on page 299, for the exceptions.
[230] One interesting thought, and a matter for future work, is making use of information on branch lengths (which appear likely to be very small if a placement is incorrect) to indicate which tree branches to try rearranging. (These rearrangements would be via flip/flopping from, e.g., ((A,B),C) to (A,(B,C)) or ((A,C),B) - A, B, and C being species in the standard tree notation - if the branch

closer to the correct tree than the starting tree was not very likely. (The number of possible trees goes up at more than an exponential rate with increasing numbers of species (Nei & Kumar 2000b).) We therefore decided to check the available literature (and with members of the committee) for possible alternatives to the starting tree, and analyze[231] the likelihoods of each manually created tree rearrangement (keeping the trees from altering in topology on an automated basis) in parallel. This was done by placing the altered trees in (otherwise PHYLIP-format (Felsenstein 1993)) files of multiple trees (with the identification being simply by what number the tree was in the file). These, along with the subsets, were processed by "nexus.add.usertree.section.pl", which also output what trees could be distinguished between by the different subsets. (It also

---

from (A,B) to C appeared to be too short.) This move has some resemblance to the existing "local" one from MrBayes (Huelsenbeck & Ronquist 2001; Huelsenbeck *et al.* 2006; Ronquist & Huelsenbeck 2003), but appears to be at least somewhat different. The parameters for this "move" would be:

- What the maximum branch length, subject to the second constraint, was that could be rearranged (possibly relative to the expected average branch length from the "brlenspr" setting; for the default "exponential(10)" setting, this is 0.1 according to the MrBayes manual);
- What the minimum number (or, possibly, minimum proportion) of internal branches possibly to rearrange is, so that this move could occur even if the first parameter were to indicate that no internal branch lengths were sufficiently short (or if too few would be usable for the possible different moves to be enough). If this parameter was used, then the internal branches chosen as possibilities would be the X shortest ones, where X would be the minimum. Another use for this parameter would be to decide how to allocate the possibilities for which branch to try rearranging, if there were very many - equal for the X smallest branches, and inversely related to the current branch length for any others.

The branch lengths used for the new tree version should, so that the move is reversible (a requirement for MrBayes' MCMC algorithm), be such that the resulting tree could be reversed back to the original if the move were to occur again. (Of course, this move should only be tried on trees for which branch lengths had been determined (either via the input of data from prior runs, or via branch length determination in the current run), not with default branch lengths. With regard to the input of data from prior runs, the validity of the method would admittedly depend on the validity (for topologically correct tree areas) of the current branch length-combining algorithm (see "Tree distances", on page 113).)

[231] To be noted is that this was done with the same randomization "seed" for each alternative tree topology; thus, differences between log probabilities are not due to chance. Similarly, the initial starting tree had arbitrary branch lengths designed not to favor any particular tree. (These were constant (0.1) for all but those inside "full" species (see "Creation of "full" species", on page 68) or between kingdoms. The former were adjusted to a lower value, while the latter were adjusted to a higher value; this process required some adjustment due to random branch length alterations

skipped any subset in which a grouping (used via a sequence or constraint) would be non-cladal, unless this was also the case with this grouping in the primary tree (number 1, used as the starting tree for the rearrangements).) For the results, please see "Tree results", on page 201.

Tree distances

One difficulty with using subsets is the question of how to deduce the overall tree distances from the subset distances[232]. Since different subsets of proteins are used for different subsets of species, while the tree distances are hopefully proportional, they are not likely to be (and, indeed, have not been found to be in the present work) anything close to identical. (For instance, a subset containing entirely mammals is likely to make more usage of, e.g., myoglobin and hemoglobin, while one containing a variety of eukaryota[233] plus some bacteria would make more usage of proteins such as ORO that have not evolved as quickly.) Indeed, this is built into our limitation to proteins that are recognizable and alignable, but have had enough evolutionary change to be of interest[234].

---

causing branch lengths to hit various internal limits.)

[232] This question arises because, due to the time and memory required, full MrBayes runs to derive distances on the entire dataset were not practical for repeated usage. (Such repeated usage would include while adjusting the proteins used by adding DHFR, altering the DHFR alignment (including by adding the deduced ancestral sequences), trying to solve the covarion problems, or fine-tuning the various other run settings (in the "props" area in particular)). It may be advisable, prior to other publications based on the tree derived, to do at least one more MrBayes run to get a better set of distances. This process may also help act as a check on how well the process described above worked, although problems with the current tree distances may also be due to some species being in fewer subsets and/or the bias mentioned below. It is probable that such a tree run will not actually be practical for all of the species; a subset with eukaryota plus a bacterial outgroup may be possible.

[233] The example is particularly applicable if the eukaryota not only included metazoa, but fungi and non-fungi/metazoa.

[234] It is possible that our limitation to proteins that have gone below a 65% identity is overly strict, in terms of both getting distances not distorted by this potential bias and the increased number of sequences available - e.g., actin for *Hartmannella cantabrigiensis*. It may be advisable to relax

Proteins that have undergone significant sequence changes inside an evolutionarily more compact set of species (e.g., mammals) are unlikely to be usable outside that set, whereas proteins that have preserved their sequence sufficiently to be alignable in a broad set of species are unlikely to have changed significantly within a more compact set of species (Halpern & Bruno 1998).

Moreover, structurally alignable portions of the proteins were run in different partitions than those that were not structurally alignable, and the latter ("nonstruct" or "uncertain") were only compared to others within the same 65% identity cluster. This may be expected to result in some differences in variability, although it is possible that the non-alignable parts vary enough faster than the structurally alignable parts to make up for the more limited distribution of them in terms of divergence - the 65% identity criterion is for the initial rough (BLOSUM80) sequence alignment of the entire protein, after all. However, such a faster rate appears only to be true *consistently* for the “uncertain” portions, if those (Brown, C J *et al.* 2002; Chen *et al.* 2006).

It is unfortunately the case that, even if one considers this proportionality, there are likely to be some distortions in distances due to the selection of proteins used and the correlations between phylogenetic closeness and commonality of protein sequences available and alignable. These distortions may be visible in the current tree in, for instance, the relatively long distances from the root for primates and the group around *C. albicans* as compared to the shorter distances

---

such a limitation for future work. Please see footnote 463, on page 267, for more discussion.

to some other species (e.g., Viridiplantae) from the root. The long distances for some species outside those concentrated on (e.g., *C. elegans*) may argue against this - on the other hand, there are considerable arguments for some species, such as many of those[235] showing long branches in the current tree, having faster mutational/evolutionary rates.

Another question in this is how group sequence distances are used; one could justify an assumed equivalence of the group position to any of:

1. the closest group species to the root;
2. the MRCA of the group; or
3. some variety of averaging of the positions of the group species (ideally weighted by how correlated the species' sequences were with the group sequence used).

Since the length of the group branches appeared too long in examined cases for the second (MRCA) option to be suitable, and the third option (a weighted average) appeared likely to take a significant amount of time to implement, the first option was chosen.

The program "estimate.starting.dists.3.pl" was written to solve the above problems. As well as the Perl interpreter, this program used two external programs, FITCH (from PHYLIP (Felsenstein 1993))[236] and a nonlinear weighted

[235] These would include Nematoda, various parasitic species such as those in the *Plasmodium* and *Cryptosporidium* genera, and species in similar circumstances such as obligate endosymbionts (Dacks *et al.* 2002; Itoh, Martin, & Nei 2002; Lartillot, Brinkmann, & Philippe 2007; Moran 1996; Wernegreen & Moran 1999; Zhu, Keithly, & Philippe 2000).

least-squares equation solver (run manually on its output). In the initial part of the program code is the following information:

1. Limits (not necessarily strictly followed, if they would conflict with other distances) on the range of distances are given. These were mostly derived using the program "figure.out.kingdom.norm.dists.pl". They were derived from the estimated maximum (for 30% or 7% identity, the latter being for a random sequence with the same proportion of amino acids) and minimum (for 65% identity or more) distances (Nei & Kumar 2000a) for either:
    a. A Poisson-correction model; or
    b. A gamma (rate variation) model with alphas of approximately 0.4 (the lowest found at the time "figure.out.kingdom.norm.dists.pl" was run), 0.65 (the Grishin alpha equivalent), or 2.4 (equivalent to an approximation of the JTT matrix).

    The numbers for 65%+ identity are also used for maximums for species that were in the same genus, since they were also entirely found within the same 65% clusters. *Candida* species are an exception for which this

---

[236] Please note that another program for deriving tree distances from an existing tree and a set of inter-species distances, with some indications of reliability (both for each distance - weighting - and in relation to the size of the distance(s) involved) making a difference in the deduction, could be substituted for FITCH. Indeed, such a substitution may be considered desirable, given that FITCH:

- is, while freely available, not completely open-source;
- is not truly intended for automated program usage (e.g., it uses fixed filenames and lacks command-line options setting);
- only takes integer weights; and
- has problems with very large weight values (crashing/halting, unfortunately without error messages or other means (e.g., "coredumps") of ascertaining the exact nature of the problem).

Note, however, that the EMBOSS (Mullan & Bleasby 2002; Rice, P, Longden, & Bleasby 2000) "EMBASSY" version of FITCH may solve at least the second of the above. However, since the EMBASSY version was only found after programs were already written to use the original version of FITCH, it was not used.

maximum is not used, due to *C. albicans* and *C. glabrata* being found in different clusters.

2. Groups of trees[237] are identified. The NEXUS-format (Maddison, Swofford, & Maddison 1997) file used for the original input for each of these groups is also identified, and a weight (estimated based on the degree of success of the "runs" in question and the overlap with groups done later[238]) given.
3. From the groups, an initial ("orig") tree with distances, which was initially created with arbitrary distances (see footnote 231, on page 111) and thereafter was the result of the prior run of the program, is identified to be used as a starting point.

The steps performed in the program's operation (some of which are repeated at least twice; see below) are:

1. The desired tree topology was read in, and species were expanded to "full" species (see "Creation of “full” species", on page 68).
2. The NEXUS-format file for each group was read in.
3. The informational entropy content of the sequences used for each prior tree determination (for each group, from the NEXUS file) was approximated[239] for each "full" species. These were used to weight the contribution from this group for the species in question; see below.[240]

---

[237] Each tree in a group was derived from identical sequence data but different runs (or different "burnin" values used to evaluate the results of the same run)

[238] Earlier group weights were reduced when later groups overlapped, and groups were removed once sufficient new data from new groups (provided said new groups had adequate-quality runs) was added.

[239] The most significant approximation involved was an equal frequency of amino acids.

[240] One possible improvement in this (besides the earlier-noted approximation) would be to note the charsets (see below) from which the informational entropy came, and for each pairing of species only using the informational entropy from charsets they had in common. I.e., the

4. The trees were read in from each group; species that had been expanded into "full" species since the time of that group's run were treated as "groups" containing the "full" species in question.
5. The species used to create each group (see "Further sequence processing: Group sequence creation", on page 96) were read in from the NEXUS file for each group.
6. Subsets of the desired tree matching the species and species group(s) found in each group were created, with arbitrary distances (as per footnote 231, on page 111).
7. The desired tree subsets were compared[241] to the topology of the trees in each group, by checking clades (groups of species descended from a common ancestor) for differences[242]; species in groups were downweighted in significance in later work in proportion to the degree of difference from the desired tree (subsets).
8. Distances for trees inside groups with multiple trees were then scaled so that the tree length (total of the tree's distances) for each tree in a group was equal to the median of the tree lengths in the group.
9. For each group, each pair of species was checked. The median[243] was found of the distances for each of the trees and the minimum and

---

minimum information entropy for each species for each charset in common would be added together. (“Charsets” are subsets of the protein groups, which were originally the established partitions until after any needed mergers (see "Partitions: Gamma, Invariant, Rate", on page 105, and "Partitions: State frequencies", on page 107); e.g., the "struct" residues for ORO would be a charset.)

[241] This step is needed because of the principle that makes distance methods work as a means of tree topology in the first place - distances between species imply a topology.

[242] This check uses the "symmetric difference" (Penny & Hendy 1985).

[243] In the present research, if there are an even number of values from which getting a median is

maximum desired distance for that pair of species. This median for the group was noted as the initial estimate (from that group) for the distance between the species.

10. A weighted[244], trimmed[245] mean was taken of the distances between species from each group.
11. For each tree, each pair of species was checked; if its distance was within the minimum and maximum desired, then the ratio of the distance found in 10 (above) and this distance was taken and put into a weighted[246] geometric[247] mean. The new between-species distances from that tree were then gotten by multiplying the resulting ratio times the old value (with some constraining for the minimum/maximum desired).
12. The new distances for each group's trees were then combined via a median (with the addition of the minimum/maximum desired if the group in

---

desired, then which of the middle two was taken as the median was determined by either a trimmed (removing the top and bottom values) mean (if there are at least 6 values) or by a normal mean. (If the (trimmed) mean was between the middle two values, then it was considered the median.)

[244] The weights are by the earlier group weights and species entropy weights (see item 3, on page 117); it is probable that an error was made that the weights from the clade comparisons (see item 7, on page 118) were not included in calculating the weights used.

[245] Trimming was only done if there were 5 or more distances. In the trimming, the middle 50% of the *weighted* values were used. Please note that this means, for instance, that if the weights for the distances 1,2,2.5,3,4 were 0.2,0.15,0.15,0.3,0.2, then the weighted mean would be of the following:

- 2 with a weight of 0.1
- 2.5 with a weight of 0.15
- 3 with a weight of 0.25

This procedure can be thought of as placing the distances on a line with lengths proportional to their weights (not to their values), cutting off the bottom 25% and top 25% of the line, and getting the weights from the new lengths for each of the distances. If the above procedure would yield having only one distance entering into the trimmed mean, then the next lower and higher distances were also entered into the mean, with their full weights.

[246] The weight used for this depended on whether the species in question were actually "full" species with the same "real" species; if not, then the effective weight was increased. It is possible that the informational entropy of the species' sequences should have been taken into account in this as well.

[247] The geometric mean is most suitable for an average of ratios; it was implemented by taking a

question had a low weight) into new estimates (from that group) of the distances in question. If this was within the minimum/maximum desired, then the ratio between the distances found in 10 (on page 119) and these distances was taken and put into a weighted[248] geometric mean. The new between-species/groups distances from that group were then gotten by multiplying the resulting ratio times the old value (with some constraining for the minimum/maximum desired).

13. The distances between species from each group were then averaged together using a weighted[249], trimmed mean[250]. The ratios between the distances found in 10 (on page 119) and these distances were taken and put into a weighted geometric mean; the new between-species distances were then gotten by multiplying the resulting ratio times the original value. The total number of sources and total weight (adjusted downward if the mean had been affected by the minimum/maximum desired) was noted, as was the weighted variance[251].

(weighted) arithmetic mean of the logarithms of the ratios (Spencer 1999).

[248] The weight used for this depended on whether the species in question were actually "full" species with the same "real" species; if not, then the effective weight was increased. It is possible that the informational entropy of the species' sequences should have been taken into account in this as well.

[249] All weights previously mentioned (group, entropy, clade conformance, and "full"/not) were used in this.

[250] In some cases, this mean was adjusted using the desired minimum/maximum (mainly by discarding one, or possibly more, values that were outside this range).

[251] There are a number of methods to derive a weighted variance, and none is truly agreed upon (Gatz & Smith 1995; Heckert & Filliben 2003; Zhang, N F 2006). The method chosen (as the simplest reasonable one that reduces to a normal (N-1) variance if all weights are identical) is to add up the weighted squared deviations from the mean, then divide by (total weight*((number of values - 1)/(number of values))). Note also that the weighted variance used all distances used for the weighted mean, not just those used for the *trimmed* weighted mean; the same is true of the total weight noted above.

14. For the first repeat of running the program (set by setting the variable "$do_stats_output" to 1), if
    a. the total number of sources was above 1;
    b. the variance was not extremely low (indicating a lack of "moves" altering the distance(s) in question); and
    c. groups were used with a non-low weight (higher than that for the previous run's tree),

    then the mean, the variance divided by the number of groups used (as per a squared standard error), and the total weight were saved for output to a ".csv" file for use for input[252] to the weighted nonlinear regression program. "estimate.starting.dists.3.pl" then halted.
15. The external program was then used to fit the datapoints, with the weights, to one of two possible equations (in which "a", "P", and "y-intercept" are all coefficients to be fitted):
    a. (variance/number of groups) = a*(mean^{P}) + y-intercept
    b. (variance/number of groups) = a*(mean^{P})

    Equation "15" was attempted initially; if it gave a negative y-intercept, then equation "b" was used (since a negative y-intercept would indicate a negative variance, which does not make sense even for a zero mean distance[253]). The "P" from the equation was used for setting the "$P_first" and "$P_second" variables (the latter using the lower of "P" and 2 - see

[252] This listing was filtered by characteristics such as the number of sources and the weights if the total was too high for the external program to handle.
[253] A positive y-intercept indicates that there is some variance (squared measurement error) even when the predicted distance is 0. This situation would not be surprising; the true distance may well be slightly above 0, even if the data indicate a distance of 0. (For instance, a back mutation

number 26, on page 125) in "estimate.starting.dists.3.pl", which was rerun after the modifications with "$do_stats_output" equal to 2.

16. A version of the desired tree was constructed with the addition of a fake "root" node[254], since FITCH requires an unrooted tree (with a trifurcation at its base)[255] and the desired tree was originally rooted (with a bifurcation between Bacteria and Archaea/Eukaryota). This was used as an input "user tree" to FITCH. The "P" for FITCH[256] was set to "$P_first" from step 15, on page 121. The distances between species (multiplied by 5 to allow for greater precision) were input, together with a rounded (see footnote 236, on page 115) version of the weights, and FITCH was run (using the default Fitch-Margoliash (Fitch & Margoliash 1967) method - modified by the "P" used - not Minimum Evolution). The resulting tree was read in, the "root" node removed, and the distances divided by 5.
17. The desired tree with the initial distances was scaled to have the same tree length as the tree from FITCH; the subtrees of this were then scaled by the same ratio. The input trees were then scaled to have the same tree length as the corresponding subtree (e.g., the "orig" tree was scaled to have the same tree length as the FITCH tree).
18. The distances between species (and groups) from FITCH were gathered.

---

may have taken place).

[254] Distances to this node were set at high enough that it should not make any difference in the resulting tree (due to the usage of a “P” higher than 0 - see footnote 256, below), and it was trimmed from the FITCH output.

[255] It would be possible to use KITSCH from PHYLIP with a rooted tree, but this would make an unnecessary molecular clock assumption.

[256] In FITCH, the variance of the measurement error is assumed proportional to the "P"-th power of the mean (Felsenstein 1993).

19. A comparison was done between the FITCH tree's distances and the intended ones (those input into FITCH). Ones that were significantly different (e.g., had a difference between these of at least 5% of the lower of the FITCH and intended distances) were output.
20. Some groups had trees that had single "full" species (see "Creation of “full” species", on page 68) when there were multiple "full" species with distances known for the particular "real" species. These were effectively substituted for by the entire "full" species subtree, using the FITCH distances (scaled proportionately so that the distance from the root to the "full" species being substituted for was preserved). In other words, the distances from the other, unseen "full" species to other species in the tree were deduced that would result in the relationships between said distances being consistent with that of the "full" species that had been seen.
21. A similar process to the above took place with actual groups[257], with the group distances in the MrBayes trees treated as equivalent to distances to the component species closest to the root.
22. For each of the newly expanded trees, the ratio between each species pair distance on the FITCH tree and the species pair distance on the new tree was put into a weighted[258] geometric mean. The new between-species distances from that tree were then gotten by multiplying the resulting ratio

[257] An additional approximation involved in this is that weighting used the outgroup's entropy weight, not the entropy weight corresponding to the residues that the species in question contributed to the outgroup sequence.
[258] The weight used for this depended on whether the species in question were actually "full" species with the same "real" species; if not, then the effective weight was increased. It is possible that the informational entropy of the species' sequences should have been taken into account in this as well.

times the old value (with some constraining for the minimum/maximum desired).

23. The new distances for each group's trees were then combined via a median (with the addition of the minimum/maximum desired if the group in question had a low weight) into new estimates (from that group) of the distances in question. If this was within the minimum/maximum desired, then the ratio between the FITCH tree's distance and this distance was taken and put into a weighted[259] geometric mean. The new between-species distances for that group were then gotten by multiplying the resulting ratio times the old value (with some constraining for the minimum/maximum desired).
24. The distances between species from each group were then averaged together using a weighted[260], trimmed mean, which in some cases was adjusted using the desired minimum/maximum (mainly by discarding one, or possibly more, values that were outside this range). The ratios between the FITCH tree distances and these distances were taken and put into a weighted geometric mean; the new between-species distances were then gotten by multiplying the resulting ratio times the original value. The total number of sources and total weight (adjusted downward if the mean had been affected by the minimum/maximum desired) was noted, as was the weighted variance.

[259] The weight used for this depended on whether the species in question were actually "full" species with the same "real" species; if not, then the effective weight was increased. It is possible that the informational entropy of the species' sequences should have been taken into account in this as well.

[260] All weights previously mentioned (group, entropy, clade conformance, and "full"/not) were

25. For the second repeat of running the program (set by setting the variable "$do_stats_output" to 2), if the total number of sources was above 1, the variance was not extremely low (indicating a lack of "moves" altering the distance(s) in question), and groups were used with a non-low weight, then the mean, the variance divided by the number of groups used (as per a squared standard error), and the total weight were saved for output to a ".csv" file for use for input to the weighted nonlinear regression program (this was filtered by characteristics such as the number of sources and the weights if the total was too high for the external program to handle)
26. The external program was then used to fit the datapoints, with the weights, as previously (see part 15 above, on page 121). If the "P" from the results turned out to be (significantly - more than standard error) greater than the "$P_second" that had already been set[261], then "estimate.starting.dists.3.pl" was rerun using this as the new "$P_second"; normally, this did not happen, and a rerun of the program was not needed (and not done).
27. FITCH was then run as previously (see part 16, on page 122), except that "$P_second" was used instead of "$P_first".
28. Some adjustments at this point were necessary to the new FITCH tree's distances, partially due to some bad data from prior runs with some errors in the programming (leading to extremely long branch lengths for "full"

---

used in this.

[261] The logic of constraining the "$P_second" to no lower than the lower of 2 and the used "$P_first" was that many of the new branch lengths were effectively copied from the old results, and would thus appear to have an artificially low variance. Therefore, the lower of 2 (the default for FITCH) and the prior "$P_first" was used as a minimum.

species relative to other sequences that were actually from the same "real" species).

29. Versions of the tree, with all distances, as a complete tree and (for display purposes) with a subset of Eukaryota only, with Bacteria and Archaea as groups, were printed.

Another concern in regard to tree distances is that, while MrBayes has a mechanism to allow rates to vary proportionately to each other[262], with "moves" (see Appendix J: MrBayes review/explanation", on page 379) that alter the assumed proportions[263], this mechanism may not have been completely effective, judging by the frequently low rates of acceptance of this mechanism[264]. The idea of presetting the initial rate proportions using, for instance, cluster data[265] was contemplated, but time constraints prevented implementation of this idea.

---

[262] The assumption that proportionality is at least somewhat preserved, while indeed an assumption and one that may be incorrect (Pupko *et al.* 2002; Rodriguez-Trelles, Tarrio, & Ayala 2001), appears necessary for DHFR reconstruction purposes. Admittedly, it may be that proteins showing a greater degree of proportionality to DHFR's rate of sequence change should be weighted more in the branch length determination for usage with DHFR; this is an area for further study. (In this regard, the usage of TS might have been valuable, as a protein that appears particularly likely (see item 2 under "Central protein candidates", on page 49) to correlate with DHFR's rate of change, albeit at a slower rate.)

[263] The default proportions were 1 (equal rates for each partition), using the "variable" setting for the rates. These were altered in some cases - see "Usage of the results of prior tree runs", on page 127.

[264] This mechanism is unfortunately not amenable to improvement via "Adaptation" (see item 6, on page 101), since it is a Dirichlet proportion alteration and not a sliding window or multiplier "move" (those are what "Adaptation" is suitable for - see "Appendix J: MrBayes review/explanation", on page 379). SA improved it somewhat in some cases, but not to a satisfactory level by the usual criterion (at least 10% of "moves" accepted).

[265] This procedure would be done using a combination of the percent identities and the existing evolutionary distances, so that proteins that appeared to be more variable (or, to be more precise for most proteins, partitions that appeared to be more variable) would have higher starting rates and vice/versa.

Usage of the results of prior tree runs

The results of earlier tree runs were used as starting parameters in two ways:

1. Using the current tree distances as the *initial* distances along any new tree (except when setting distances to semi-arbitrary values so as to enable comparisons of likelihoods between topologies; see page 112, footnote 231 above). This was done by "put.dists.on.tree.pl"[266];
2. Using, in some cases the results of earlier runs were used to deduce other parameters, such as rates, alphas for gamma rate variation, and invariant proportions. (This was limited to cases in which the subset of species and of proteins in use was very similar or (such as when wishing to extend a run for more generations) or identical.) In such instances, the "sump" command in MrBayes was used to extract the most likely values and/or the most likely ranges of values, along with indications of reliability such as "PRSF". The sump results were then interpreted by the program "sump.summarize.pl" and the output of this interpreted manually[267] and by the programs "use.mrbayes.sump.freqs.info.pl" and "use.mrbayes.sump.freqs.info.2.pl".

---

[266] This program is similar to "estimate.starting.dists.3.pl" (see "Tree distances", on page 113) in its usage of FITCH (Felsenstein 1993) to take a set of distances between species (in this case, from the full tree) and turn them into distances on another tree (in this case, a subtree, including potentially outgroup branches on this subtree).

[267] Manual interpretation was used both before the (current) programs were written and in order to make changes other than those these programs are capable of making, including:

- To narrow the range within which some parameters could vary;
- To conclude that some sequence partitions did not appear to have significant internal rate variation, as indicated by an alpha value (for gamma) that was quite high (e.g., 50+), so gamma rate variation should not be used for those areas.

5. Alignment of central sequences

Structural and initial sequence alignments

The structural sequences (of DHFR, DHFR/TS, and TS) were aligned similarly to the above procedure (for non-central sequences), except that no usage was made of existing database alignments of these, since said databases would potentially be influenced by the target (fungal) DHFR and TS structures. Therefore, the structural alignments were done locally (see “Locally created structural alignments”, on page 80), without using structural alignments done elsewhere. TS sequences were not found to be necessary[268], and thus alignments of TS sequences (other than structural ones) were not performed; the TS portion of DHFR/TS sequences (as judged by alignments to sequences with known placement of the DHFR/TS transition point) were removed. Alignments to the DHFR structures were then conducted as per "Sequence alignments", on page 88 (although with manual review (and selection of alignments to combine), unlike the above). This procedure was done using the program "align.to.central.3.pl", with manual consolidation of its output if the matrices differed in their results, until the alignment would fall below the 65% threshold[269] with regard to the (DHFR) structure(s) in question.

[268] The likely contribution of a TS alignment to the tree was considered low in relation to the difficulties of an additional "special-case" (like DHFR, requiring manual entry, partially due to its combination with DHFR in some sequences) alignment.

[269] This happened for fungi/metazoa after the addition of *Strongylocentrotus purpuratus* (purple sea urchin), which is the only non-vertebrate deuterostomate with a known DHFR sequence.

Alignment using HMM

Further (below 65% identity to a usable structure's sequence) DHFR alignments were by HMMER (Durbin *et al.* 1998; Eddy 1998; Eddy & Birney 2003) using the recode3.20comp prior (Wistrand & Sonnhammer 2005) and an "--idlevel" of[270] 0.65, followed by manual revision. In this procedure (which was/is one of the more likely places for errors to enter, given the considerable uncertainties involved[271]), a hidden Markov model (HMM) was generated using the existing alignment (similarly to how automatic alignments of further sequences to Pfam "seed" alignments are done (Bateman *et al.* 2002)). With this alignment, areas of (alignment) uncertainty[272] were considered "insert" regions (using the "--hand" option to HMMER's "hmmbuild" program and the "RF" line in the Stockholm-format sequence file). To be noted is that, with regard to phylogenetic work, all fungi/metazoa DHFRs were considered to be in the 65% identical cluster, although the actual percent identity fell below this level[273]. In addition, to be

[270] This uses a 65% identity level, as per the earlier alignment work.

[271] One reason for this uncertainty is, as noted, the need for alignment in some sections; it is difficult, to put it mildly, for a human being to keep track of an alignment with, at the end, 106 sequences (plus 17 alternative alignments, giving 123 total) and 459 positions.

[272] These ("nonstruct" or "uncertain" regions) are not treated as aligned outside a 65% cluster (or, for fungi/metazoa, the fungi/metazoa cluster, as with "nonstruct" and "uncertain" regions, as noted on page 129).

[273] It is admittedly somewhat unclear as to whether considering these to all be in the same "cluster", and thus the "nonstruct" and "uncertain" areas to be considered alignable, was a good idea. It was partially forced at the time by the way the programs were set up, which required the sequences of a cluster to be associated with a known (3D) structure. The alternatives for purposes of homology modeling would be:

- Using loop searches (see "Loop searches", on page 157) to predict the structure of non-"struct" regions;
- Using some combination of the modeled templates (e.g., Urdeuterostomia for Fungi/Metazoa) with increased usage of loop searches.

These would still leave the question of how to deduce the ancestral sequence in such areas, which would require an alignment. Trying multiple possible alignments (which has already been done to a mild degree, as noted on page 131) followed by testing via modeling (John & Sali 2003) may be a possible method. If the alignment differences were confined solely to areas of considerable uncertainty (e.g., loop regions), then this may be computationally practical.

noted is that the "fungi/metazoa" DHFR cluster included the species *Hartmannella cantabrigiensis*, identified by prior research (Stechmann & Cavalier-Smith 2003) as the closest[274] extant species to the root of fungi and metazoa.

For the construction of this HMM using HMMER's "hmmbuild" program, weighting was done using the current tree distances (Altschul, Carroll, & Lipman 1989; Felsenstein 1973, 1985b) estimated via MrBayes (see "Tree distances", on page 113). The process used was such that the existing sequences in the alignment were weighted in proportion to the likely closeness of the new sequence(s) to be aligned to them (in terms of phylogenetic closeness). For species with multiple sequences - due to polymorphism, due to uncertainties in the alignment (see below), or both - weights were allocated among them in proportion to their weights by a modified[275] Blosum weighting scheme. Weights of sequences from species not on the tree[276] were allocated based on the weights of the existing species combined with the modified Blosum weighting scheme.[277]

---

[274] More precisely, it is the closest extant species with a known DHFR sequence. Please note that some other possible candidates, such as *Corallochytrium limacisporum*, appear to lack (identifiable) DHFR sequences entirely. Among the reasons, incidentally, that the species *Hartmannella cantabrigiensis* is thought to be close to the divergence of fungi and metazoa is that it has seperate (indeed, on opposite coding strands) DHFR and TS sequences; this is otherwise a characteristic only of definite fungi and metazoa among eukaryotes. (Stechmann & Cavalier-Smith 2003)

[275] The modifications consisted of ignoring "insert" regions (as specified via “--hand”) and not increasing the weight of sequences due to gaps. The modified version is invoked via “--wpb2”.

[276] Several *Plasmodium* species (*gallinaceum*, *vinckei*, *inui*, and *cynomolgi*) were not included on the tree due to lack of other sequence data and the difficulties in assigning their proper location with respect to other species (except with regard to their well-established identities as in the *Plasmodium* genus; they are used for laboratory work on malaria).

[277] Please see the programs "find.species.weights.3.pl", "transfer.weights.to.stockholm.1.pl", and the (local, except for some material from prior research (Wistrand & Sonnhammer 2005) that was not used due to being more for searches than for alignments (Wistrand 2005)) source code modifications to HMMER's "hmmbuild.c" at patchfile “hmmbuild.c.patch” for exact details (of this

Following alignment with HMMER's "hmmalign" program[278], manual modifications were made as necessary; this was primarily[279] needed because areas considered "insert" regions[280] are not aligned by HMMER, except to place them between appropriate non-insert regions. This non-alignment caused some degree of uncertainty[281] in several areas; as a result, it was decided to enter some sequences twice (with two different alignments).

The alignment proceeded progressively, with a HMM being built using an existing alignment, weighted according to the phylogenetic position of the target species or group of species (e.g., *C. briggsae* and *C. elegans* were aligned at the same time). The already-existing portion of the alignment was held constant for the

---

somewhat elaborate procedure) if desired. (Another modification was the addition of the "--nofrag" flag, used to communicate to the program that sequences were never to be treated as fragmentary.) The "--wme" maximum-entropy weighting scheme was initially tried, but it unfortunately assigned a weight of zero to some species; it was concluded that this method was more suitable for usage with HMM building for searches than for alignments, since it is ultimately mainly intended to maximize the discrimination between matching and nonmatching sequences. (The creation of a weighting scheme designed to maximize the discrimination between "good" and "bad" alignments is an interesting idea for future work.) The alterations will be submitted to HMMer's authors after some minor output formatting issues are cleaned up.

[278] The critical option used was "--mapali", to use the current (used for hmmbuild) alignment as an alignment to which the new sequences were to be aligned.

[279] Another case was with sequences of markedly greater or lesser length. In some cases, these could be determined to be due to an alternative start site; the material prior to the site analogous to the human/mouse/chicken known-structure sequences was removed. In other cases, this was due to the sequence being a fused DHFR/TS gene (as with eukaryota other than fungi and metazoa); the linker and TS portions were trimmed off, insofar as they could be distinguished (such as via prior analysis of the sequences in SWISS-PROT (Boeckmann *et al.* 2003)). The *Ustilago maydis* sequence has a notably long end extension, with an uncertain alignment of other (fungal) sequences versus this extension; please see "Appendix K: Partial DHFR alignment", on page 384. Examination of the DNA sequence (not performed locally yet) may indicate a sequencing error causing the non-recognition of a stop site.

[280] These are indicated in the Stockholm-format files by a "." instead of an "X" in the RF line.

[281] To be noted is that the manual alignments made use of, when possible, information from structures outside a given cluster, despite earlier-identified uncertainties as to the correspondence of some residues between clusters. It is possible that this attempted identification of residues was mistaken, and/or that too little importance was placed on minimizing gaps (as opposed to similar residues being found in a given column).

automatic portion of the alignment (the HMM aligned the new sequences to the existing alignment, via HMMER's "--mapali" option to "hmmalign"), but could be revised manually in light of the new sequences. The alignment was also revised as the modeling progressed in light of structural findings, as well as predicted ancestral sequences being added to the alignment as follows:

1. All DHFR sequences except for some fungal and predicted ancestral sequences were in the alignment prior to modeling. The fungal sequences not present were Ascomycota other than *P. carinii*.
2. Each predicted ancestral sequence (or set of sequences) was added to the alignment after the creation of the models for it/them (the model creation enabled the elimination of some of the originally postulated possible sequences - see "8. Examination of models", on page 352).
3. After the Uramniota and later predicted sequences were added to the alignment and new determination of tree distances, a HMM was generated that was weighted by closeness to the goal species, *P. carinii* and *C. albicans*. A consensus sequence was derived from this HMM by HMMER's "hmmemit" program using the "-c" option. A `blastp` search (using the settings for the BLOSUM80 searches run previously) was done to see which Ascomycota were the closest to this consensus sequence; those found by the search and considered (by visual inspection of the alignment in the `blastp` output) to be alignable were selected to be aligned next[282].

[282] For instance, after adding the Uramniota sequences, the *Yarrowia lipolytica* sequence was added; the *S. pombe* DHFR sequence was not chosen for alignment at this point, despite its phylogenetic closeness to the already-aligned sequences, due to its considerable number of insertions.

6. Determination of ancestral sequences

The object of ancestral sequence determination is finding the most likely sequence at a given ancestral node[283]. The determination of ancestral sequences can be divided into two parts, the determination of the amino acid (or possible amino acids) at a given position and the determination of whether an ancestral sequence position has a residue present or not (i.e., non-gap versus gap)[284]. For the first two nodes for which (DHFR) ancestral sequences were determined, namely Urplacental[285] (the ancestor of placental mammals) and Uramniota (the ancestor of amniotes[286]), no gap determination was considered necessary, insofar as:

- There are no structurally known[287] gap differences among placental organisms; and
- The sole difference for the *G. gallus* structure was found to be at the end, an uncertain area in any event.

For further ancestral sequence determinations, it was necessary to determine gaps; this determination is probably more error-prone than the amino acid predictions (it is less well understood, as discussed in item 3 on page 38). In

[283] More precisely, one is finding (in most cases) the most likely *set* of ancestral sequences, due to uncertainty in reconstruction and the possibility of polymorphism.
[284] The latter is of (primary) concern for areas where present-day sequences descended from the node in question differ with respect to insertions and deletions (gaps).
[285] Except for the fungi/metazoa common ancestor, the ancestral nodes will be named as "Ur" (meaning "primordial") followed by the most applicable taxonomic identifier for the node (followed by a per-sequence identifier if more than one probable ancestral sequence was found). See Figure 3.4, on page 149, for a diagram of the locations of the ancestral sequences determined.
[286] For our DHFR dataset, amniotes were mammals plus *Gallus gallus* (chickens).
[287] Some - generally uncertain - sequences (such as the two for *Canis lupus*) without known (3D) structures had gaps relative to other placental organisms.

such cases, gap determination took place prior to sequence determination, both as a matter of logic (it would be difficult to determine what residue was present when there was no residue present!) and because some residues are more or less likely to appear in the vicinity of gaps (Chang, M S S & Benner 2004).

In terms of file names, *etc.*, note that most are named according to the associated sequence variation (e.g., KD for a lysine (K) in one position and an aspartic acid (D) in another position). For stages with gaps present (Urdeuterostomia and after), this is either modified by the insertion of underscores (e.g., _E_SKFEDQ, abbreviated "ES") or prefaced by a numeric code followed by an underscore (e.g., 0011_KD), with more 1s generally[288] indicating more gaps. In writing about groups of these, the wildcard character “*” may be used (e.g., 0011_K* would include both 0011_KD and 0011_KP).

[288] Due to residue and gap correlations noted (see page 138), in some locations a "1" indicates a gap in one position and a "0" indicates a gap in another position. Gap coding also can affect the residues seen at other positions, and similarly residues in different positions may be indicated by one letter if they are correlated.

Sequence determination

For sequence determinations involving gaps (all after Urplacental and Uramniota), the program used (after determination of gaps - see "Gap determination", on page 139) was "nexus.extract.ancestral.seqs.3.pl"[289]. This program[290]:

1. Read in the original sequences, their phylogenetic groupings, any fixed state frequencies, and the initial values for any Dirichlet state frequencies (see "Partitions: State frequencies", on page 107) from the NEXUS file used to run MrBayes for the ancestral sequence determination.
2. Read in what the possible gap arrangements ("ids") were;
3. Determined, for each position that was present (non-gap) in any of the gap ids[291]:
    a. All of the amino acids seen at that position ("all_orig_AA").
    b. If there were any species with amino acids at that position that were either close by the target (e.g., for Urascomycota, fungi), had known structures, or had modeled structures, the amino acids seen in those species ("orig_AA")[292]; otherwise, "orig_AA" was identical to "all_orig_AA".

[289] The programmatic procedure for sequences not involving gaps was similar, but simpler.
[290] Please note that the below description is of the latest version of the program; some changes have been made to, for instance, adapt the program to differing subsets of species of interest for different ancestral sequences desired.
[291] For "nonstruct" and "uncertain" positions, the sequence examination was restricted to species in the fungi/metazoa cluster.
[292] Note that, in this case, “orig_AA” is a subset of “all_orig_AA”.

4. Read in the sump results (see under “Usage of the results of prior tree runs”, on page 127), as summarized by "sump.summarize2.pl"[293].
5. Extracted any summarized Dirichlet state frequencies (see "Partitions: State frequencies", on page 107) from the sump results.
6. Examined each position in the sump results, determining the range of probabilities of the amino acids at that position, and:
    a. Determined whether the position was "near a gap" - in an insertion, directly next to an insertion, or directly next to a deletion for each gap id (if gap ids differed on this, the position was considered to be neither near a gap nor *not* near a gap).
    b. If the position was "near a gap", and did not have as one of its “orig_AA”s one or more of the gap-associated residues "D", "G", "P", or "S" (Chang, M S S & Benner 2004)[294], but one of these was one of its “all_orig_AA”s, then the gap-associated residue was added as an "orig_AA".
    c. If the position was not "near a gap", and did not have as one of its “orig_AA”s one or more of the non-gap-associated residues ("F", "Y", "W", "M", "I", "L", or "V"), but one of these was one of its

[293] The range (minimum-maximum) was determined by sump.summarize2.pl by examining the likely degree of validity of the position, according to the PSRF (below 5 being considered good), degree of variation, and other characteristics. If it appeared valid, then the minimum was the lower of the mean and median, and the maximum was the higher of the mean and median. If it did not, then the minimum was the lower of the mean and the 5th percentile of the results, and the maximum was the higher of the mean and the 95th percentile of the results.

[294] Note that this set of residues, and the “non-gap-associated” residues also used, are a more conservative sets than the full ones listed by the paper referenced; the modifications took into account similarities based on the ESIMILARITY matrix (see “Appendix G: ESIMILARITY matrix”, on page 374).

"all_orig_AA"s, then the non-gap-associated residue was added as an "orig_AA".

7. For each residue among the "all_orig_AA"s that had a significant maximum probability (at least 0.05; higher (up to 0.1) was required of extremely common residues, according to the frequencies), the following were classified as possible residues:
    a. If it was among the "orig_AA"s.
    b. If it had a significant minimum probability (at least 0.075; higher for extremely common residues), and one of:
        1. was a glycine
        2. was a proline (Visiers, Braunheim, & Weinstein 2000; Yang, W Z *et al.* 1998)
        3. among the "orig_AA"s was a glycine
        4. among the "orig_AA"s was a proline
        5. the maximum probability was very high (at least 0.2; higher for extremely common residues)
8. For each gap id, gave an output of positions versus amino acids considered possible at that position[295] and associated sequences (one with only residues that were 95%+ probable present, another with residues that were over 50% probable present (the "chars2" sequence); these had "x"es for positions of uncertainty). The output of positions, including gaps, was

[295] This included the probabilities; these were the average of the minimum and maximum, scaled to 1 total for all residues considered possible at that position.

done in respect to[296] the earlier stage(s) that were to be used as the templates for modeling.

Unfortunately, the above procedure tended to produce too many possible sequences to model (especially with a semi-manual procedure; see "7. Model building", on page 146). The primary means to reduce this excess was by looking at correlations between residues in the existing aligned sequences, using the programs[297] "find.residue.correlations.pl" and "find.residue.correlations.2.pl" with manual interpretation. This procedure also included examining correlations between residues and the presence of gaps; however, this was extremely difficult to work with on a manual basis given, for instance, the size of the DHFR alignment (and its problems with gaps; see "5. Alignment of central sequences", on page 336). This procedure could result in the elimination of some possibilities; if, for instance, a reasonably-certain residue was a "G" and having a "G" at that position was associated with *not* having an "A" at another, uncertain position, this helped narrow down the possibilities at the second position. It could also result in the conclusion that the identities of two uncertain residues were correlated (e.g.,

[296] While this format was the most useful for the immediate task of modeling, it was unfortunate in some cases with regard to placing the new sequences into the alignment with inserted residues in the proper position (in regard to what position the ancestral sequence determination was actually using). On the other hand, this lead to some corrections of the alignment.

[297] These used chi-square with accounting for multiple comparisons, plus some examination of potential structural correlations via the output of "find.interacting.res.pl". The latter's success was difficult to determine, due to some errors in earlier versions of “find.residue.correlations.pl” (the only program used for earlier stages). These bugs prevented any of the structurally close residues from being used at stages prior to their correction, while at later points much of the uncertainty examined was with regard to insertions/deletions and backbone-affecting residues (glycine/proline differences). These would be expected to affect amino acids not within H-bonding (3.8 Ang. max (Kahn 2007c)) or (for charged atoms/residues) reasonably possible (9 Ang.) Coulomb interaction distances, due to the potential for the entire backbone flexing and consequent significant movement of distant residues.

"DK" or "SY"); this was one cause for the examination of multiple possible sequences.

Gap determination

For purposes of ancestral gap determination, positions[298] were coded (by the program "nexus.add.gap.partitions.pl") in two ways:

- Binary, as suggested in the MrBayes user manual (Huelsenbeck *et al.* 2006), with a 0 for a non-gap and a 1 for a gap. This data was placed into a "restriction" partition in MrBayes, with a coding type (allowing for that no all-gap positions would be seen) of "nopresencesites"[299] and a rate variation type of gamma[300]. For positions with polymorphism with regard to gaps[301], the '?' ambiguity symbol was used.
- Into the following categories, coded as if DNA:
  - A: Not a gap
  - C: 1-3 residue gap (including this position)
  - G: 4-7 residue gap
  - T: 8+ residue gap

[298] Note that gaps are not being used as contiguous units, but per position (whether an amino acid is present or not), although the second means of coding gaps (described on page 139) takes into account the *surrounding* length of gaps. This methodology was chosen because of the difficulties with defining gaps in terms of insertions/deletions when one does not know (yet) what the amino acid arrangement was in the ancestral sequence undergoing said insertions/deletions.
[299] Some error messages concerning the coding were seen (incompatible data positions for "coding=nopresencesites"), for reasons that (due to time limits) have yet to be determined. These did not prevent the program from running properly as far as could be determined (for instance, no locations appeared to be missing in the predicted ancestral gap probabilities).
[300] The "gamma" rate type was used since neither invgamma (invariant + gamma) nor adgamma (discussed on page 140) are allowed for the binary/restriction datatype.
[301] These were primarily for (synthetic) outgroup sequences - see "Further sequence processing: Group sequence creation", on page 96.

These numbers were determined based on an earlier study (Goonesekere & Lee 2004), with the breakpoint of 3 being suggested in that study and that of 8 being determined by local examination of graphs from said earlier study. (For polymorphic positions with regard to gaps, ambiguity codes (using brackets in the NEXUS format), allowing for more than one possibility, were used.) Substitution probabilities for this were determined by a GTR[302], with initial values approximated from the earlier study (Goonesekere & Lee 2004). Two types of rate variation were tried (in two partitions), invgamma (invariant plus gamma) and adgamma (gamma with correlations between adjacent positions, since (for instance) a T in the above coding is certain to be found with at least one T next to it[303]). The GTR transitions and state frequencies for these partitions (one for invgamma and one for adgamma) were linked.

For both of the above, partition state frequencies were done as a Dirichlet (with the starting Dirichlet frequencies determined by the gap proportions seen).

Two methods were chosen because each has advantages and disadvantages:

- The binary coding method has been used previously, as noted, and is relatively simple (and thus, e.g., is less vulnerable to programming errors and may be less vulnerable to overparameterization). However, it does not take into account the general finding that two one-residue gaps are less common than one two-residue gap; insertions/deletions can insert or delete a group of

[302] This allows a substitution matrix to be estimated from the data (and the assumption of time reversibility), although a starting point is needed.

[303] It is unfortunately the case that this is not a particularly good means of simulating said correlation; how to do so better is an open problem, as with most areas of handling correlations between positions.

residues at once (indeed, for a frameshift not to occur, insertions/deletions require at least 3 nucleotides to be inserted/deleted). It also does not allow for using rate variation methods other than “gamma”.

- The second (“DNA”) method takes into account gap lengths, but has not (to our knowledge) been previously explored and has some complexities (e.g., finding the proper transition matrix between states) associated with it.

For comparison purposes (which appear to require a simpler dataset - see “Discussion and future work”, on page 344), and to try to use the combined advantages of each by using both as much as possible, both methods were tried. We contemplated using the “standard” datatype from MrBayes, by which the exact gap length could be coded, but:

- This would not take into account the distributions previously found of the lengths of gaps (Goonesekere & Lee 2004), since MrBayes does not estimate substitution rates for “standard”-coded data and, for determining character frequencies, treats each site in the “standard” model as independent (making overparameterization likely);
- This would, again since MrBayes does not estimate substitution rates for “standard”-coded data, either:
  - By using “unordered” for the “ctype” setting, not take into account that a gap is more likely to transition between two lengths that are closer together (e.g., from 10 to 12) than it is to transition between lengths that are further apart (e.g., from 0 to 10 or from 1 to 20); or

  - By using “ordered” for the “ctype” setting, force the assumption that all transitions between gap lengths took place only one at a time (potentially distorting the tree distances and topology).
- This would be more useful if we were coding for gaps as insertions/deletions, not positions (although the same could be said of the “DNA” coding method); and
- This appeared likely to lead to over-weighting the importance of the gap data (particularly inappropriate given the uncertainties of gap models in evolutionary biology!) in terms of, for instance, distance determinations (including those done implicitly during the ancestral sequence determination runs).

Also contemplated was dividing the gap lengths into 20 categories (including “no gap”) and coding as amino acids, but:

- The resultant need for MrBayes to determine a substitution matrix appeared likely to overparameterize the model (Huelsenbeck *et al.* 2006);
- Again, this would be more useful if coding for gaps as insertions/deletions, not positions;
- Again, this appeared likely to lead to over-weighting the gap data.

One difficulty seen with regard to gap coding was for "nonstruct" or "uncertain" areas (with the latter being particularly common) outside the fungi/metazoa cluster. Initially, these were coded into the gap partitions also, but this lead to considerable difficulties in interpreting the results for the Urdeuterostomia

(vertebrata plus sea urchin, in our case) common ancestor concerning gaps - for this set of sequences, only the "binary" gaps were used. Subsequently, only the common (reliably aligned) portions of the DHFR sequence, plus the "nonstruct" or "uncertain" portions in the fungi/metazoa cluster, were used, with a question mark (for missing data) used for the latter portions in non-fungi/metazoa species in general[304].

The decision was also made that "nonstruct" and/or "uncertain" portions of the sequences would not be in a separate partition in MrBayes for gap determination, despite that they appear likely to show differing characteristics[305]. This decision was due to worries about having a sufficient amount of data, particularly for the "DNA"-coded version given the usage of a GTR, plus the disruption to the "DNA"-coding mechanism and interpretation of it from such a split (and the disruption of the adgamma correlations). Later work should check, among other matters (see "Discussion and future work", on page 344) whether at least the "binary"-coded partition could/should be split up.

The program "nexus.extract.ancestral.seqs.2.pl" performed the initial estimation of gap positions; its function was analogous to "nexus.extract.ancestral.seqs.3.pl"

[304] The exception was that, if it could be determined what the equivalent position would have to be in terms of gap. For instance, if no residues were present between known-equivalent ("struct") residues in the *P. falciparum* sequence, then this area could be considered all-gap. If this is unclear, please see the program for exactly how this worked.

[305] This likelihood is most evident for the “uncertain” portions, since they were selected in the first place due to their gaps.

in its initial stages (1 (on page 135) and 2 (on page 135) to 4 (on page 136))[306]. After these initial tasks, the stages were as follows:

1. The deduction of positions where it appeared reasonably certain that the desired sequence was a gap or not a gap, based on the presence or absence of the gap in all "close" sequences (analogous to the "orig_AA" for "nexus.extract.ancestral.seqs.3.pl").
2. Thresholds (minima and maxima) were found for what probabilities would be accepted as strongly saying "gap" or "non-gap", from both binary and "DNA"-coded gap sequences (with the latter being a comparison of "A" versus "C", "G", and "T" together). Please see "Gap determination thresholds", on page 342, for more information on the background of this stage. The thresholds are set so that the correct proportion (by the state frequencies[307]) of the positions would be expected at the next stage to be gaps[308] by said thresholds.
3. The thresholds determined earlier were compared to the minimum and maximum probabilities (of a "0" or of an "A", respectively) to determine whether a position appeared likely, on an initial basis, to be a gap or a non-gap (or not determinable - '?'), and the degree of certainty of this evaluation. This process was done separately for the "binary" and "DNA"-coded gap information.

---

[306] One matter that was noted was that the probabilities for "A", "C", "G", and "T" for the "DNA"-coded positions did not always add up to 1 (probably due to a high degree of uncertainty causing a wide range between the minimum and maximum predicted). Cases significantly (greater than +/- 0.01) deviating from 1.0 total were treated later as potentially uncertain.

[307] Note that the desired proportions in question were a range, using, first, both the original and sump-determined frequencies, and, second, both codings of gap data.

[308] Note that this also used the evidence from stage 1 - positions already determined to be

4. Two potential gap/non-gap sequences were then created, "binary" priority and "DNA" priority. Each of these used its "priority" if it was clear, or if the other did not contradict.
5. Three "DNA"-format sequences were created using data from stage 1 and stage 4 for whether positions were gaps, non-gap, or undetermined.
6. What these sequences would allow in terms of the "DNA" formatted gaps[309] was compared to various means of determining the validity of some means of telling (by simple probabilities, by minimum/maximum ranges, *etc.*) how reliable a possible "DNA"-format prediction was[310].
7. The rules deduced from the above were used manually[311] to modify the program's next stage. This stage used them to predict what the "DNA"-format sequences should be (including what the likely level of validity of positions should be), with separate sequences (and, if necessary, rules) for "invgamma" and "adgamma".

The "DNA"-format sequences output at the end were then put together with the sequence resulting from the "binary" coding to help resolve uncertainties and construct a set of possible gap arrangements (with, for instance, evidence of a "T" being most probable used to indicate that an uncertain-length gap was at least 8 residues long). While automated determination of correlations was

---

reasonably certain to be gap or non-gap were classified accordingly, without thresholds.

[309] E.g., if there was a possible 1-residue gap surrounded by non-gap positions, then "G" and "T" were not allowed.

[310] I.e., seeing if it contradicted earlier conclusions if one simply used the highest-probability letter, with cases of yes/no classified by the validity determination means.

[311] Much, even most, of the rules in question unfortunately appear currently to change with each sequence predicted, probably partially due to alignment problems and resultant inadequate degrees of correspondence between positions in various sequences; further research, preferably with a more reliable alignment (see “Discussion and future work”, on page 344), is desired.

attempted, as with sequence determination, this was not very successful[312]. Gap determination ultimately necessitated a considerable amount of manual examination[313] of the alignment.

The possible gap sequences were then fed into the program "nexus.extract.ancestral.seq.pos.pl", along with the original NEXUS file used for the ancestral sequence determination run(s); the result was fed into "nexus.extract.ancestral.seqs.3.pl" (see "Sequence determination", on page 135).

7. Model building

Two possible choices were available with regard to homology modeling:

1. The usage of already-available automated modeling programs such as Modeller (Fiser, Do, & Sali 2000; John & Sali 2003; Sali & Blundell 1993; Sali & Overington 1994; Sali *et al.* 1995; Sali 2001; Sanchez & Sali 1997a, 1997b) or SWISS-MODEL (Schwede *et al.* 2003). With this option, ancestral structures would be modeled at, perhaps, each node on the tree (or, rather, at least each time that the predicted ancestral sequence changed at all). This frequency of modeling would be partially to make up for the increased inaccuracy in automated modeling techniques at lower sequence identity levels (Bowie, Luthy, & Eisenberg 1991; Mosimann,

[312] Please see under “Sequence determination”, on page 138, for more on the difficulties involved in this.
[313] This process of manual examination did contribute to the realizations made of problems with said alignment - a helpful, if at the time depressing, result.

Meleshko, & James 1995; Sanchez & Sali 1997b; Saqi, Russell, & Sternberg 1998; Taylor 1994 Dalton, 2007 #3214).

2. The usage of more general (and, ideally, open-source) programs (e.g., GROMACS (Berendsen, van der Spoel, & van Drunen 1995; Lindahl, Hess, & van der Spoel 2001; Lindahl *et al.* 2007; van der Spoel *et al.* 2005) for energy minimization, possibly with restraints (Flohil, Vriend, & Berendsen 2002; Sali & Blundell 1993; Sali 1995)) with manual modifications[314] as necessary. With this option, ancestral sequences would be simulated for fewer nodes (preferably, only for those with gap or other potentially critical changes), since the labor of repeated homology modeling would be significantly greater while the accuracy of the modeling would (hopefully) be improved over that in automated means.

The second option above was chosen (see Figure 3.4, on page 149, for the phylogenetic positions of the structural models created), for several reasons including:

1. The lack of availability of open-source homology modeling programs (e.g., Modeller does not come with source code, and neither is it redistributable, especially with changes). This lack of availability makes it additionally desirable to write open-source programs, or at least programs on which more-automated modeling software could be based[315].

[314] Such manual modifications include the local authorship of new programs and alteration of old ones as necessary.

[315] Note that the programs in question do not, in general, require visual examination of protein structures; this is of assistance in automation considering the limited abilities of current computer vision algorithms (although the use of an algorithm created for computer vision was used in the creation of the “Nussinov” matrix (Naor *et al.* 1996)). In this regard, it is perhaps fortunate that a committee member's examination of the usefulness or lack thereof of visual examination of

2. That many programs were likely to have included fungal DHFRs in the databases used to create them (e.g., for Modeller's potentials[316], or for loop searches using SWISS-MODEL (Schwede *et al.* 2003)); it is unclear how to restrict programs to ignore this information, or if it is even possible (this appears unlikely for Modeller's potentials, for instance).
3. That most modeling programs, on the other hand, would not take as full advantage as may be possible (e.g., see footnote 337, on page 157) of the non-fungi/metazoa known DHFR structures (from *Plasmodium* and *Cryptosporidium* species), since they are too far away to serve as practical templates.

It unfortunately appears likely that, while the second option was indeed the correct one, too large jumps in terms of sequence identity (particularly for a largely self-taught homology modeler like the author) were chosen. On the other hand, at most stages multiple models (not only in terms of modeled sequences, but also in terms of modeled structures) were nonetheless created, which one might not anticipate being possible without a fully developed automated homology modeling software package. It should be noted that some elements of the software in question are still under development; the below is only regarding those elements that were used in the research thus far.

---

models for their evaluation came to a negative conclusion as to its utility, despite the member's original belief otherwise (Kahn 2007e).

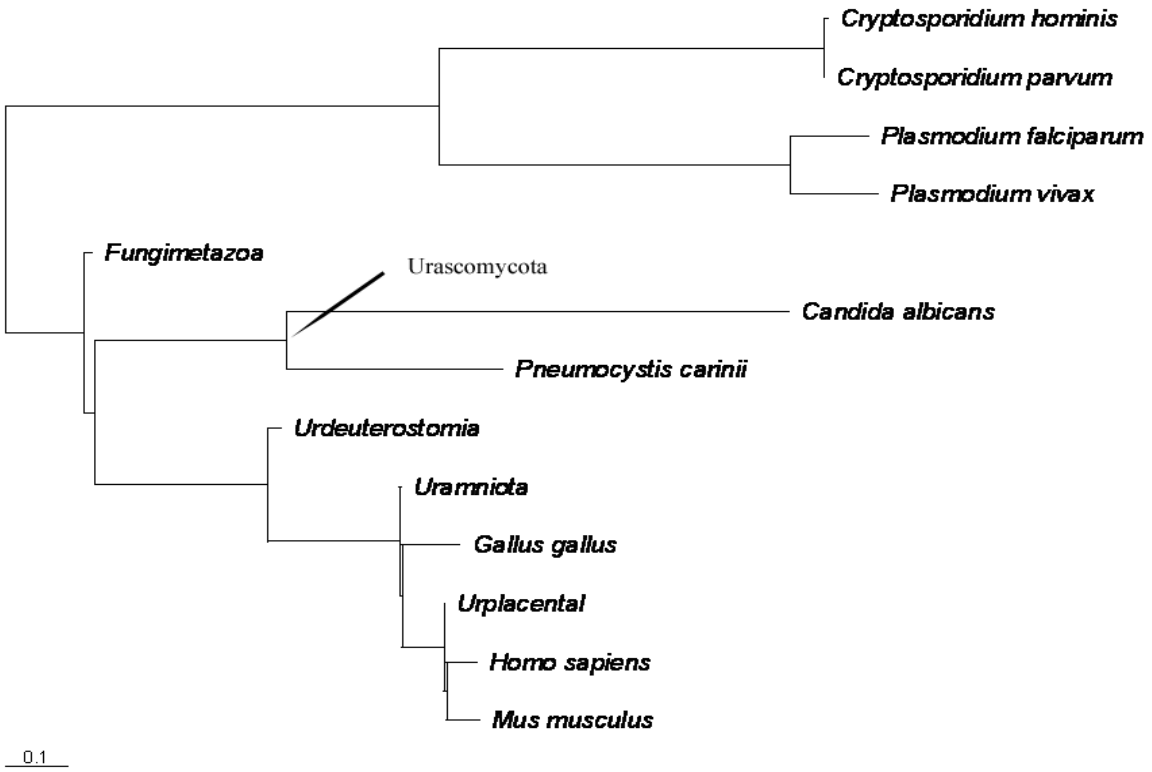


Figure 3.4: Positions of structures and models

Another matter to note is that ".mdp" files are locally created files that instruct[317] GROMACS' "grompp" program in how GROMACS' "mdrun" program[318] is to behave. The mentioned ".mdp" files are available under http://cesario.rutgers.edu/easmith/research/mdp/ and in the “mdp.tar” supplemental file (in UNIX “tar” format).

---

[316] Admittedly, it is possible that such programmatic usage of fungal DHFRs, at least with regard to "potentials" and not loop searches, would not actually be a sufficient portion of the databases used as to be a problem. It was, however, felt - especially given the other considerations - that it would be better to avoid any potential bias. (Note also that some structural databases, particularly the older ones, used for various previous studies have been quite small (e.g., 500 or less) due to the need for stringent cutoffs for quality (Lovell *et al.* 1999, 2000; Richardson, D C & Richardson 2001). Such stringent cutoffs are particularly needed for X-ray crystallographic and NMR work since otherwise errors from existing structures will be amplified, and modeling frequently makes use of methods built on such databases (e.g., MolProbity - see "MolProbity", on page 186).)

[317] Readers may wish to consult the GROMACS online reference to their syntax at “http://www.gromacs.org/documentation/reference/online/mdp_opt.html”. Note also that GROMACS uses nm, not Ang.; the same is thus true of much of the following discussion.

[318] The “mdrun” program does energy minimization and simulated annealing.

GROMACS was compiled with single precision numbers (instead of double precision, which would likely cause a significant slowdown and/or significantly greater memory consumption and is generally not recommended for energy minimization and simulated annealing) using the FFTW (Frigo & Johnson 2005) FFT package. Hydrogens were added when necessary by `reduce` (Word *et al.* 1999a; Word *et al.* 1999b; Word & Richardson 2006) run with a dot density of 100 per square Ang. as an improvement on the default 16 (per square Ang.).

Assignment of initial coordinates

Initial coordinate assignment used an approximate (see footnote 327, on page 150) geometric median of the coordinate positions in the templates. The programs to do this are still under development, but for the most recently completed models, the program "put.together.pdbs.sequence.3.pl" was used. The stages were as follows:

1. Input files[319] containing information on:
    a. What sequence was desired
    b. What the template files were, and for each template:
        1. What residues were considered "bad"[320] for their mainchain atoms[321] (see "MolProbity", on page 186)
        2. What residues were considered "bad" for their sidechain atoms

[319] These can be found in the supplemental file “put.together.pdbs.tar” (in UNIX “tar” format) and at http://cesario.rutgers.edu/easmith/research/put.together.pdbs/.
[320] Residues considered “bad” were still used, but to a lesser degree - see item 6, on page 153.
[321] One revision on this, incorporated into current work, also considers "bad" the (backbone) carbonyl carbon and oxygen of the residue prior to "bad" residues and the backbone nitrogen and (for non-proline) hydrogen after "bad" residues. Another revision automatically considers "bad" atoms in residues next to an insertion or deletion, unless the insertion or deletion is present for all

3. If the template should be considered "old"[322]; these were considered "bad" for everything except locations considered "bad" by all non-"old" templates

c. Which template to align the other templates to[323], if this had not already been done (see below). This indicates the "align to" template.

d. What the alignment of the templates to the model sequence was

2. Any residues considered "bad" for mainchain atoms for all templates were warned about (and no longer considered "bad"), and similarly for sidechains.

3. If necessary, the templates were aligned to the "align to" template, so that the coordinate systems were on a common frame of reference. (This alignment used only the residues that would be used for the model, not any that would be deleted.) Copies of these aligned files were saved so that they could be put into the input file for any further runs of "put.together.pdbs.sequence.3.pl", avoiding the need for future alignments.

---

available templates.

[322] “Old" here refers to models two stages removed from the model being created (e.g., the Urdeuterostomia models were considered “old” for creating the Urascomycota models, while the fungi/metazoa ancestral models were not). Among the reasons this was done was that sometimes the predicted ancestral sequence "flip/flopped" on which residue was predicted at a location (e.g., it might be "T" for Urdeuterostomia, "A" for the fungi/metazoa ancestor, and "T" again for Urascomycota), and failing to use the next level up (Urdeuterostomia for Urascomycota, in this example) would lose information (e.g., the proper location of the sidechain OH and CH3). (It is not meant as an insult to those older than the author, especially given the author's (growing) awareness of the author's aging.) A similar process, with "groups" of models, was used in some earlier stages - one group would have some templates considered the equivalent of "old" and the other group would have other templates considered the equivalent of "old", to keep parallel models going.

[323] This was normally whichever template had the fewest mainchain "bad" (if this was tied, then the MolProbity evaluation was consulted, using the highest "good" phi/psi angles then the lowest "bad' bond angles).

Alignments were done via lsqrms (see "Locally created structural alignments", on page 80), using the "Identity" matrix if no sequence differences were present, and the normal set of matrices if otherwise.

4. The desired and template file sequences were examined to determine, for each position, what residue was needed and what residues were available. Residues that differed from the desired one were automatically considered "bad" for sidechains, and would be considered "bad" for mainchain atoms if the difference were of glycine versus non-glycine[324] or proline versus non-proline.
5. In some cases, even though a residue was not the same residue as desired, by "trimming off" (ignoring) some atoms, positions[325] could be derived. (For instance, by removing the sidechain OH, the positions of atoms in tyrosine could be converted to the positions of atoms in phenylalanine. For a full listing of these, please see under "Loop searches", on page 159.) Moreover, even when this was not possible

[324] It is possible that the program should check, for each glycine to non-glycine alteration, on whether each template's phi/psi angles were such that a non-glycine could fit said angles (Lovell *et al.* 2003). If so, then (for that template) simply due to a substitution being from glycine to non-glycine should not mean that it was "bad". Whether the other way around (non-glycine to glycine) should be considered non-"bad" is questionable:

- While one could theoretically check for whether each template had problematic phi/psi angles for the existing (non-glycine) residue as an indicator of how glycine might be used in such a situation, the usage of residues that may be otherwise strained (to try to accommodate such angles) is questionable.
- The case may differ depending on whether the alteration is going backward or forward in evolutionary terms:
  - if a residue is being mutated from a non-glycine to a glycine, then this may be an indicator that the new residue needs unusual phi/psi angles
  - if a residue is being mutated from a glycine to a non-glycine, then evidently (barring major conformational changes) the original phi/psi angles should be reasonable.

This question may also bear onto the case of "glycine to non-glycine" transitions, above.

[325] Admittedly, in some (perhaps most) cases, these positions would be approximations in regard to having the proper distances - while, e.g., the sidechain oxygen in serine and the sidechain sulfur in cysteine correspond, they are not the same atom by any means. (Going from, for

exactly, in many cases one of a selection of atoms[326] could be considered equivalent (e.g., one of the sidechain hydrogens for alanine to the sidechain OH in serine); if one had other information for which atom should be chosen (a difficulty in many cases), this could be used. Similarly, it used when possible even a partial set of atoms that were equivalent (e.g., phenylalanine when the desired residue was tyrosine), if the full set was available. A set of coordinates from the various sources, for each desired residue's atoms, was assembled with this analysis.

6. An approximation[327] of a geometric median was performed, with the following steps, to derive each atom's position:
    a. For each source atom, if the atom was one that could be said to be relative in position to another atom[328], and the position of the other atom had been determined[329], then the source atom's effective XYZ position was adjusted so that it was in the same place relative to the other atom in terms of its XYZ coordinates. (For instance, if the source had a alpha carbon at 0,0,0 and a beta carbon at 1,1,1, and the current model has a alpha carbon at 1,0,1, then the source would

---

instance, isoleucine to valine is another matter.)

[326] This procedure, while potentially valuable, was the source of some errors, such as atoms winding up coinciding with each other in position due to using the same other atom as a source. One example of this was when attempting to predict the positions of asparagine's sidechain from aspartic acid's sidechain, or vice-versa. It is uncertain to which (equivalent) oxygen in aspartic acid is the nitrogen in the asparagine sidechain equivalent. A future version of the program may do this via looking at hydrogen-bonding and/or steric hindrance, similarly to `reduce`; this would also be useful for determining which hydrogen in alanine was equivalent to the OH of serine.

[327] A geometric median is unfortunately a challenge; the method used is an approximation algorithm (Weiszfeld 1937). It should, however, achieve the goal of avoiding outliers having too much influence (via down-weighting them), as would a true geometric median (similarly to an arithmetic median).

[328] For instance, the position of a beta carbon would be relative to the position of its alpha carbon.

[329] The other atom's location should have been determined already, given the order in which the

be considered to contribute a beta carbon at 2,1,2.) It is unfortunately probable that this should have instead been done with regard to angular relationships (using more than one atom to base these on), since the XYZ orientation of residues is essentially arbitrary - this is a matter for future work.

b. If there are multiple atoms from a source corresponding to the desired atom, the other sources are averaged (with equal weights) and the closest atom is taken. If there are no other sources, then for each source all atoms are averaged (with equal weights), then the sources are averaged (with equal weights), to determine the starting position. Another round of averaging then takes place, weighting by the inverse of the distance to the starting position (equivalently to the below 3 steps (c-e); please read those first if this is unclear) for each atom within a source, with each source then weighted equally. The resulting position is used to determine which atom to use (the closest one for each source).
c. The atoms from each source are averaged together (with equal weights) to determine a rough position.
d. The distances from the atoms from each source to the current rough position are determined.
e. The atom positions from each source are averaged together again, with weighting equal to 1 over the distance from the previous step, yielding the second rough position.

---

search was done.

f. The distances from the atoms from each source to the second rough position are determined.

g. The atom positions from non-"bad" sources alone are averaged together, with weighting equal to 1 over the distance from the previous step, yielding the final position.

h. An approximation of the crystallographic "temperature"[330] was derived from the (unweighted) RMSD of the final coordinates versus the sources, by squaring it (i.e., getting a Mean Square Deviation) and multiplying by $8*pi^2$ (Rhodes 2000). If this was above[331] 40, and the atom was a hydrogen, it was skipped (hopefully to be later added by `reduce` in a more appropriate position).

7. Deletions and insertions were handled in this program by, respectively, leaving residues out and loop searches (see "Loop searches", on page 157).

Following the above, `reduce` was run on the resulting model files, and any hydrogens it removed (due to steric clashes) were put back by "restore.reduce3.removed.pl" (partially to enable better functioning by GROMACS).

---

[330] Among the reasons for doing this were to pass the information to other modeling programs (see, e.g., under "Loop searches", on page 157) and, potentially, to display programs (showing which atoms were the most variable (at least in respect to the templates)).

[331] 40 is the default crystallographic temperature at or above which waters are ignored by, for instance, `reduce`.

NADPH insertion

The alignment and insertion into the DHFR structure of NADPH[332] was done[333] by the program "average.hetatm.4.pl", which:

1. Took each of the source (PDB) files (with NADPH present) and aligned their amino acid chains[334] (via "align.lsqrms.wrapper.full.pl"[335]) to the (PDB-format) file that needed NADPH inserted;
2. Read in what the locations of the atoms of the NADPH were;
3. Performed an approximation of a geometric median, as follows:
   a. The (xyz coordinates of the) points were first averaged (via a mean) together, with weights proportional to the alignment quality (see footnote 167 on page 82) for the source file.
   b. The points were then averaged again, but with weights inversely proportional to their distance to the results of the first average; the result was the coordinates used.

---

[332] It was chosen to insert NADPH since most of the DHFR structures available, including the sources and targets, have NADPH bound, so its presence should help to constrain the modeling to a realistic conformation (e.g., able to have NADPH bound). The other substrate, dihydrofolate (or folate, in some instances) was unfortunately not present (due to stability and/or enzyme action) in most structures, with the usual substitute (if any) being an inhibitor. Therefore, neither dihydrofolate nor folate ligands were modeled. No inhibitor was modeled due to the differing properties of DHFRs from different species with regard to inhibition (Appleman *et al.* 1988a; Appleman *et al.* 1988b; Baccanari *et al.* 1989; Blakley & Sorrentino 1998; Brophy *et al.* 2000; Degan *et al.* 1989; Farnum *et al.* 1991; Lewis *et al.* 1995; Shallom *et al.* 1999; Taira & Benkovic 1988).

[333] In the first few rounds, this procedure was carried out (partially) manually.

[334] It is probable that the alignment should have used only those residues closest to the NADPH (or those that should be closest, using an input alignment). It is possible that such an alignment should have used atoms other than the main-chain heavy ones, although this would require knowing, for each residue, which atoms were interacting with the NADPH. A check on the quality of the NADPHes in the models, probably only possible visually, is desirable to see how critical these are for future work.

[335] As implied by the name, this program carries out an automatic Structal alignment (see "Locally created structural alignments", on page 80), including the output of an aligned structure. Indicated by "full" in the name is that it tries matrices other than the Identity one.

4. The (unweighted) RMSD of the new coordinates relative to the source points was found, and translated into a "temperature" (see "Assignment of initial coordinates", on page 150). If the atom had a "temperature" over 40, and was a hydrogen that had been added by `reduce`, then it was skipped (since `reduce` should be able to add it again, in a better location). Otherwise, the coordinates, with the temperature, were put into the output file along with the original contents of the file needing NADPH inserted.

This process was carried out prior to any minimization in all rounds after the first (Urplacental); in the first, it was done after vacuum minimization.

Loop searches

Due to the existence of some sequences, including lengthy ones, that fold into very different structures despite being identical (Jacoboni *et al.* 2000; Zhao *et al.* 2001), loop searches generally look primarily or exclusively for conformance of the geometry of the anchor points (residues already in the model on each side), including their orientation with respect to each other. This method of search from our viewpoint had several problems:

1. It ignores, at least in its simplest form, that some residues are likely to have distinct effects on the geometry of a loop. This is particularly true of proline and sometimes glycine but also, for instance, those involved in intra-loop charge interactions or tendencies toward alpha helix formation[336].

[336] This consideration is for a sufficiently long loop. Note that beta sheets are generally not on the exterior of proteins (Richardson, J S & Richardson 2002), the usual location for loops. Extended strands are an exception, but these are usually recognizable by their tendency toward a high proline content (Eswar, Ramakrishnan, & Srinivasan 2003).

2. It may ignore when the source for a loop is, upon doing a sequence search, obvious, namely loops in the correct location in another homologous structure[337] - admittedly, for our purposes this had to be restricted to avoid usage of the target structures by accident.
3. It ignores the possibility of using the loop as a source of information on rotamers[338], thus helping avoid[339] rotamer searches (see "Rotamer searches", on page 164) for such conformationally flexible side chains as that of lysine[340].
4. Perhaps most importantly under the circumstances, implementing it would necessitate the creation of a database of protein geometry. It is quite possible that said database would need to be created newly (locally) due to licensing restrictions, concerns about contamination with target structures, *etc.*, with a resultant significant expenditure of time (on a heavily mathematical area not particularly suitable for being handled by Perl, the main programming language used in this project's *de novo* software).

---

[337] This happened with the loop search to correct problems in 13-27 in Urdeuterostomia, for which loop searches were not otherwise run; the search turned up the *P. falcip.* and *Cryptosporidium* DHFR structures, which were those that turned out to align properly in the area in question. (This region is positions 47-63 in the alignment shown in “Appendix K: Partial DHFR alignment”, on page 384. It overlaps with the region of the predicted sequences in Figure 1.1, on page 4.)

[338] Insofar as rotamers are dictated by, for instance, the surrounding amino acids and insofar as their validity may be correlated with how well the loop will fit into the protein (Chakrabarti & Pal 2001). Note that by “rotamer” is also meant the location of the beta carbon, when appropriate.

[339] It was unfortunately suggested, and the suggestion followed, not to do loop searches for Urdeuterostomia, with the exception of one area (13-27), due to the belief that rotamer searches and (for insertions/deletions) heuristic means (e.g., averages and extrapolations of locations) would do the job adequately. The resulting time to put together the rotamer search programs, and the time to run the rotamer searches themselves, caused a considerable loss of time, especially insofar as the loop search programs were needed eventually in any event for more lengthy insertions (more than 1 residue).

[340] In the rotamer library used (Lovell *et al.* 2000), lysine has 27 groupings of rotamer angle *ranges* seen (see footnote 352, on page 164) with the most common three having only 20%, 13%, and 6% of the instances known. The library in question was constructed using only residues for which all side chain atoms could be located with confidence. Arginine is similarly problematic.

Loop searches, when an area of sequence had been identified as problematic (because of an insertion, because of a change in glycines/prolines, or because of the addition of a difficult-to-rotamer-search residue like lysine), were performed via scanprosite (de Castro *et al.* 2006) on the ExPASy server[341]. The patterns used[342] had four major sources:

1. Those using the beginning or end of the chain as an anchor, due to the desired sequence being at the beginning or end of the PDB file.
2. Those based on a combination of:
    a. from what residues another residue's rotamer[343] could be predicted (e.g., the rotamer for "S" can be predicted from that for "C", at least in terms of steric hindrance; see item 5 under "Assignment of initial coordinates", on page 152); and
    b. the ESIMILARITY matrix (see "Appendix G: ESIMILARITY matrix", on page 374).

    As one might guess, these groupings were particularly used when avoidance of a rotamer search was considered particularly valuable. This

[341] However, scanprosite can be downloaded. In hindsight, it would probably have been best to do so, despite the need also to download more PDB file sequences for it to scan (since the local listing of PDB file ATOM-record sequences is limited to PDB chains thought to potentially be of use for other reasons). Such downloads and local setup would be a necessity for further automation (see “Loop searches”, on page 348).

[342] In usage, the patterns were substituted for the desired residues in the section of sequence needing a loop search. For the initial search, the most specific pattern (other than only the desired amino acid) was used for residues other than the anchor residues (at the ends), for which the least specific pattern was used. If this generated too few (one structurally known sequence or no structurally known sequences) or too many (above about 100, due to the need for manual processing), then the patterns were adjusted for greater or lesser specificity or the length was varied.

[343] If any - no side chain information is needed for glycine. Again, “rotamer” includes the beta carbon location (e.g., alanine does not have a rotamer, but does need information on its beta carbon location).

was created manually; the groups[344] are in order of increasing preference (specificity):

Desired Residue	scanprosite Coded Groups				
A	(G)	(GP)	[ACS]	A	
C	(GV)	(GVP)	C		
D	[DFYWNLH]	[DN]	D		
E	[EQ]	E			
F	[FY]	F			
G	x	[GSN]	G		
H	[HW]	H			
I	I				
K	[KR]	K			
L	[LFYWNDH]	[LF]	L		
M	[MQKR]	M			
N	[NHW]	[NH]	N		
P	P				
Q	Q				
R	R				
T	[IT]	T			
S	(GV)	(GVP)	[TSNKQED]	[TSN]	S
V	[VIT]	[VI]	V		
W	W				
Y	Y				

[344] These are in the appropriate coded form for input to scanprosite, except that (due to limits imposed by reference management software in use) parentheses should be read as curly brackets (indicating all residues *except* those listed are allowed). (For the others, an "x" indicates any residue, a set of brackets indicates allowed amino acids (e.g., histidine or tryptophan for "[HW]"), and a single letter indicates that amino acid only.)

3. Those meant for insertions that were long enough that they might have a helix in them, or that were in the area of a helix and so might start or end it. The sources for these groups were the (central residue) helical propensities from EMBOSS' (Rice, P, Longden, & Bleasby 2000) garnier.c (Garnier, Osguthorpe, & Robson 1978), together with ESIMILARITY (see "Appendix G: ESIMILARITY matrix", on page 374), and the "Nussinov" matrix. These were put together by "find.helix.coil.sub.groups.pl" to produce (again, in order of preference/specificity, coded for scanprosite):

Desired Residue	**scanprosite Coded Groups**			
A	[ACDEKLMNQRSTWY]	[AEKLMQW]	A	A
C	[ACFILMVWY]	[CIY]	[CI]	[CI]
D	[ADEHKNQRST]	[DNRST]	[DNRST]	[DNRST]
E	[ADEHKNQRST]	[AEHKQ]	[EHKQ]	[EHKQ]
F	[CFILMVWY]	[FLMVW]	[FLMVW]	[FLMVW]
G	[ACDEGHKNQRST]	[CDGNRST]	G	G
H	[DEHKNQRSTY]	[EHKQ]	[EHKQ]	[EHKQ]
I	[CFILMVWY]	[CIY]	[CIY]	[CIY]
K	[ADEHKNQRST]	[AEHKQ]	[EHKQ]	[EHKQ]
L	[ACFILMVWY]	[AFLMVW]	[FLMVW]	[FLMVW]
M	[ACFILMVWY]	[AFLMVW]	[FLMVW]	[FLMVW]
N	[ADEHKNQRST]	[DNRST]	[DNRST]	[DNRST]
P	[ADEHKNPQRST]	[DNPRST]	P	P
Q	[ADEHKNQRST]	[AEHKQ]	[EHKQ]	[EHKQ]
R	[ADEHKNQRST]	[DNRST]	[DNRST]	[DNRST]
S	[ADEHKNQRST]	[DNRST]	[DNRST]	[DNRST]
T	[ADEHKNQRSTWY]	[DNRSTY]	[DNRST]	[DNRST]
V	[CFILMVWY]	[FLMVW]	[FLMVW]	[FLMVW]
W	[ACFILMTVWY]	[AFLMVW]	[FLMVW]	[FLMVW]
Y	[ACFHILMTVWY]	[CITY]	[IY]	[IY]

4. Those meant for short insertions, without worries about rotamers (or for which the rotamer search set was found to be too restrictive), that might be inside beta-sheets or at their edges (as well as the perhaps more likely helix or coil). The sources for these groups 7were the ESIMILARITY (see "Appendix G: ESIMILARITY matrix", on page 374) and the "Nussinov"

matrix[345]. They were put together by "find.sub.groups.pl" (again, with groups in order of preference/specificity, with groups coded for scanprosite):

Desired Residue	scanprosite Coded Groups		
A	[ACS]	[AC]	[AC]
C	[ACFILMV]	[ACFILMV]	[AC]
D	[DEHKNQRST]	[DEHKNQRST]	[DENQS]
E	[DEHKNQRST]	[DEHKNQRST]	[DEHKNQRS]
F	[CFILMVWY]	[CFILMVWY]	[FILMWY]
G	[GNS]	G	G
H	[DEHKNQRSY]	[DEHKNQRS]	[EHNQR]
I	[CFILMVWY]	[CFILMVWY]	[FILMV]
K	[DEHKNQRST]	[DEHKNQRST]	[EKNQRS]
L	[CFILMVWY]	[CFILMVWY]	[FILMV]
M	[CFILMVWY]	[CFILMVWY]	[FILMV]
N	[DEHKNQRST]	[DEHKNQRST]	[DEHKNQRST]
P	P	P	P
Q	[DEHKNQRST]	[DEHKNQRST]	[DEHKNQRS]
R	[DEHKNQRST]	[DEHKNQRST]	[EHKNQR]
S	[ADEHKNQRST]	[DEHKNQRST]	[DEKNQST]
T	[DEKNQRST]	[DEKNQRST]	[NST]
V	[CFILMVWY]	[CFILMVWY]	[ILMV]
W	[FILMVWY]	[FILMVWY]	[FWY]
Y	[FHILMVWY]	[FILMVWY]	[FWY]

The initial program to process the search results was "extract.needed.pdbs.for.loop.search.pl"; it extracted which PDB files needed to be downloaded[346] (if any), and checked for PDB file chains mentioned in multiple places in the file.

The next program to make use of the results of these searches was "put.together.pdbs.section.3.pl", which was followed by

[345] Given the derivation of the ESIMILARITY matrix (see "Appendix G: ESIMILARITY matrix", on page 374), this essentially means that the major contributor to these groups was the Nussinov matrix, with BLOSUM62 mainly mattering when the Nussinov matrix was unclear.

[346] Note that one problem encountered was that recent PDB files are in a different format (v3) that we do not use; fortunately, the download source (ftp.rcsb.org) does not include any v3 files, and they could thus be recognized by their absence.

"put.together.pdbs.section.4.pl". These worked similarly to "average.hetatm.4.pl" (see "NADPH insertion", on page 156), but:

1. Instead of the initial alignment being of an entire PDB file, they took a section[347] out of another PDB file and aligned it to the residues of the model that were supposed to correspond (the "anchor" residues), according to the input data file.
2. The quality of the alignment from item 1 was used instead of the quality of the overall alignment for weighting (the "temperature" from the original file was used to determine the "quality" of its atoms - see under "Assignment of initial coordinates", on page 155).

They were also similar to "put.together.pdbs.sequence.3.pl" in three regards:

1. A third round of averaging (approximating a geometric median) was done; some residue sources were classified as "bad" and only used for the first 2 rounds. The decision on which residues were "bad" depended on the main chain for "put.together.pdbs.section.3.pl", while it depended on the side chain for "put.together.pdbs.section.4.pl". The former concentrated on places where there was an insertion or a difference in glycine versus proline versus others (e.g., it called the original file "bad" if this was true there). The latter concentrated on places where there was a difference in sidechains (and the new sidechain could not be taken from the old sidechain - see item 5 under "Assignment of initial coordinates", on page 152).

[347] As well as using the exact section specified in its input file, it could also use a partial version with some of the anchor residues trimmed off; the highest-quality (discussed on page 82) version

2. Both programs technically[348] had the capability to use sidechains of other residues to get information on the proper positioning of the desired residue's sidechain atoms.
3. Both used positions relative to other atoms when applicable; e.g., the backbone N and C positions for each source residue were done in XYZ coordinates[349] relation to the position of the alpha carbon for that residue, and thus they were in the same position relative to the new position of the alpha carbon.

The input files for both programs can be found in the supplemental file "put.together.pdbs.tar" (in UNIX "tar" format) and under http://cesario.rutgers.edu/easmith/research/put.together.pdbs/.

Rotamer searches

Rotamer searches were primarily done via the programs "prekin"[350], "mkrotscr", and "`probe`"[351] (Word *et al.* 2000), using a previously published rotamer library (Lovell *et al.* 2000) in its fuller, online[352] form. The first uses of these programs

---

tried from a particular PDB chain for a target residue was used.

[348] This capability, in general, was probably a mistake for put.together.pdbs.section.3.pl, in hindsight; it should have left this to put.together.pdbs.section.4.pl. In addition, as with "put.together.pdbs.sequence.3.pl" (see "Assignment of initial coordinates", on page 150), there were problems encountered with this resulting in atoms overlapping/coinciding in position.

[349] As previously noted (see "Assignment of initial coordinates", on page 150), this was probably an error; angular (i.e., a spherical coordinate system) relationships are more appropriate.

[350] For any future use of this, we recommend the non-GUI version, which was unfortunately not available at the time of downloading (which is probably responsible for some - possibly all - of the program crashes that have been seen).

[351] As with `reduce`, a setting of 100 dots per square Ang. was used for `probe`, as an improvement on the default 16.

[352] The online version (Richardson, D C & Richardson 2001) used has asymmetric bin widths. For purposes of determining rotamers of Asp and Asn, the secondary structure was assumed to be non-alpha-helix, non-beta-sheet ("coil").

and the library, for the Urplacental sequence's residue[353] 127, were manual. The second round of uses, for the Uramniota, were more automated, but the process still took a significant amount of time[354] (partially for programming the process of interpreting the initial search results and shrinking the ranges used (so that they could be searched in better detail), partially for waiting on the rotamer searches themselves). In the Urdeuterostomia and later stages, the programs used for these searches were "create.mkrotscr.mutate.1.loop.pl" and "create.mkrotscr.mutate.2.loop.pl".

Rotamer searches were also used to compensate for earlier problems with the initial coordinate assignment (see "Assignment of initial coordinates", footnote 326, on page 153). These errors were extremely close atoms - in the worst cases in the exact same position; they were detected by the program "check.for.bumps.2.pl".

Translations to/from GROMACS, PDB formats

GROMACS has a built-in mechanism to translate from PDB to GROMACS format, the program "pdb2gmx". However, partially due to the somewhat inconsistent atom naming nomenclature in the PDB (at least for the v2.3 file format, which is that used in this study), particularly for non-proteins such as

[353] Unfortunately, the predicted sequence had this residue changed to a lysine; see footnote 340, under “Loop searches”, on page 158, for more about why this was unfortunate.
[354] The considerable amount of space taken up by the output should also be noted.

NADPH, some modifications were necessary, including to the "xlateat.dat"[355] GROMACS file. Besides changing this file, a program was created ("reformat.pdb.gromacs.pl") that did some other alterations. These alterations are based both on the "xlateat.dat" file and on the "force field"[356] of interest (set in the header of the program). In particular, the program checks on cases where GROMACS uses several four-letter (or longer) residue/heteroatom (non-protein) names to distinguish between different states[357] for which the PDB format uses only one, three-letter code. It distinguishes between these by which one corresponds best to the atoms seen in the PDB file, and puts in the new residue name. This program is frequently called by another, "to.gromacs.wrapper.2.pl", which:

1. If necessary, runs `reduce` to add hydrogens,
2. then, if `reduce` was run, "restore.reduce3.removed.pl" to restore any pre-existing hydrogens removed by `reduce` due to clashes[358],
3. then "reformat.pdb.gromacs.pl",
4. then "pdb2gmx"[359],

[355] This file acts to translate between PDB and GROMACS atom nomenclatures. For a listing of changes, please see the patchfile "xlateat.dat.patchfile". Some of the changes in question, but not all, were suggested by a posting on the GROMACS mailing list (de Groot 2004).

[356] In molecular dynamics (used in simulated annealing) and energy minimization, a "force field" is a file - or set of files - containing information on how to simulate a chemical structure and its movements. The simulation is largely using classical (not quantum) mechanics (since quantum mechanics is far too difficult to compute for anything large); the simplifications involved in using classical mechanics can be described as the most critical part of a force field. The force fields in GROMACS also including information about the "topology" of chemicals (how the atoms are connected together) and how to add hydrogens to them if said hydrogens were not already present in the input PDB (or similar) file.

[357] The most important ones for our purposes are the charge/protonation state of histidine and NADP(H).

[358] While it is preferable for `reduce` to determine a better hydrogen atom location if possible, if it is unable to do so, then we chose to retain the original position to avoid discarding information when better (deduced) location data are not available.

5. then "add.restraints.wrapper.pl" (see "Creation of restraints", on page 170).

Similarly, translations back from GROMACS to PDB format can involve some difficulties with nomenclature. The program "reformat.gromacs.pdb.pl" acts to translate the more troublesome differences. The program "do.reduce3.restore.pl" can then be used to run `reduce`[360] and "restore.reduce3.removed.pl" to restore any hydrogens removed.

Partially frozen vacuum/dry minimization

The initial translation into GROMACS format (see "Translations to/from GROMACS, PDB formats", on page 165) for all vacuum/dry minimizations was with a modified[361] GROMOS96 43b1[362] (vacuum) force field (van Gunsteren *et al.* 1996). If any insertions or deletions were done, then a vacuum (dry) energy minimization with most atoms frozen was done, to try to relieve local problems (if not, then the next stage was the creation of restraints - see "Creation of restraints", on page 170).

---

[359] This program was run with the appropriate flags to dictate the force field in use and to tell GROMACS to use an H-bonding distance of 0.38 nm (Kahn 2007c).

[360] This stage is necessary because the force fields thus far used in this research do not include some hydrogens explicitly; they are instead accounted for in adjusted atomic sizes, *etc.* (The hydrogens omitted are those bound to carbons, generally, with the exception that those bound to aromatic ring carbons are included.)

[361] The modifications were to add some hydrogens to the NADPH that were either not added by `reduce` or did not translate properly to GROMACS; see patchfile "ffG43b1.hdb.patch" for the modifications.

[362] Note that the 43b1 force field explicitly includes all hydrogens except those attached to aliphatic (non-aromatic) carbons; hydrogens attached to aliphatic carbons are subsumed into the mass, charge, and other properties of these carbons.

The program "create.ins.del.mutate.freezegrps.2.pl" was used to determine what atoms should not be frozen. There were two major categories of "unfreezing":

1. In some cases, only one full residue was unfrozen:
    a. If a residue was mutated, then at least that residue, the (backbone) carbonyl carbon and oxygen prior to it, and the backbone nitrogen and (for non-proline) hydrogen after it were unfrozen;
    b. If atoms were involved with clashes that prevented `reduce` from adding hydrogens (see "Translations to/from GROMACS, PDB formats", on page 165), then those atoms (including the hydrogen, if GROMACS added it) were unfrozen;
2. In other cases, 3 residues[363] before (and the (backbone) carbonyl carbon and oxygen prior to that) and 3 residues after (and the backbone nitrogen and (for non-proline) hydrogen after that) a particular residue were unfrozen, as well as the residue itself. This happened if:
    a. The residue was inserted;
    b. The residue was next to a deletion;
    c. The residue was mutated to or from a glycine[364] or proline; or
    d. The residue was involved in a clash causing problems for `reduce` in adding hydrogens (see "Translations to/from GROMACS, PDB formats", on page 165).

[363] It was not possible to do 5 residues before/after - this would encompass most of the protein, making "freezing" likely to be useless - and therefore 3 residues were used (Kahn 2007b).
[364] It is possible that alterations from a glycine not adopting unusual phi/psi angles (Lovell *et al.* 2003) to a non-proline should not have been treated as unusual.

Vacuum minimization was done without the use of a "box" (see footnote 379, on page 175), without Coulomb interactions (using an infinite effective dielectric), but with the ranges of the VdW cutoffs chosen to encompass the entire molecule for their upper limits (which is only possible without a "box" or with an extremely large "box"). As is recommended by GROMACS (to avoid a "tumbling ice cube" with simulations without a "box"), all motions of the center of mass of the system were cancelled. The energies of interactions between the frozen atoms were excluded. The stages of vacuum energy minimization were as follows:

1. The first stage used the mdp file "vacuum.with.ins.del.mutate.freeze.mdp". This file calls for an energy minimization: using the "steepest descents" minimizer[365] (a slow but very reliable minimizer, the latter characteristic being why it was chosen). The minimization terminated when either the maximum force (on any one atom) dropped below 500 kJ/(mol*nm) - this value was a guess - or roundoff limits (for movement of atoms or reductions in energy) were reached; normally, the first happened.
2. The second stage used the mdp file "vacuum.with.ins.del.mutate.freeze2.mdp"; this file calls for the same energy minimization as the above, except that it was using the "conjugate gradients" minimizer (which is faster but sometimes less reliable when far away from the energy minimum). The energy minimization terminated

[365] Please see footnote 71 under “7. Model building”, on page 40, for a review of the role of the minimizer.

when either the maximum force went to below 89 kJ/(mol*nm)[366] or roundoff limits were reached; which happened first was variable.

Creation of restraints

Unlike in many homology modeling efforts, the present work has the potential advantage of having aligned, homologous structures that, while too far away to serve as templates, may serve as a source of data in other ways. The idea - a heavily modified form of previous work (Sali & Blundell 1993; Sali 1995) - was to analyze the various available structures for the distance relationships between corresponding atoms, and use the NMR distance restraints[367] code in GROMACS to restrain the models to follow any patterns spotted. In general, the ranges were found by examining the known DHFR structures, then allowing for some level of variability outside the range of these[368]; what level of variability was allowed depended on whether the restraints were "non-strict" (more variability allowed) or "strict" (less variability allowed).

---

[366] This was the approximate (depending on the model) number of residues that were fully non-frozen.

[367] The distance restraints code in GROMACS includes, advantageously, the ability to set both minimum and maximum distances expected. It also has the capability to use more than two atoms in a given restraint, although because the GROMACS code is meant for NMR work this can become complex (see "Simulated annealing when needed", on page 183, regarding the distance restraints used in that work for one example).

[368] The major influences on the allowed degree of variability were the range of distances seen between DHFRs from different genera and the range of distances found for multiple DHFR structures from the same genus.

The distance restraints were of two types:

1. Restraints on distances between particular atoms of the NADPH[369] and atoms of binding or related positioning[370] residues in the DHFR. This was determined as per the above brief description, with it being restricted to DHFRs with NADPH bound. This procedure was performed by "find.AO.NO.AOP.constraints.2.pl", using an input file that gave the residue correspondences[371], which group of atoms in the NADPH the restraints were with, and whether the relationship was with the main-chain or (more commonly) the side-chain of the residues. This program then output a file that was processed by "create.restraints.2.pl" along with a "topology" file created by GROMACS (describing the identities of the atoms), to create a set of restraint values. These were output into another file, which was included in GROMACS' processing by appropriate statements in the "define" line of the ".mdp" file in use. (In other words, different ".mdp" files

[369] These were, for the version used for Urascomycota, the charged oxygens bonded to the pyrophosphate linkage (and not to other atoms), the adenine phosphate oxygens, and the nicotinamide 6-membered (5 carbons and 1 nitrogen) ring. The addition of further atoms may be desirable, such as those determining the "handedness"/stereochemistry (A/B specificity) of binding and thus, for instance, from which side of the NADPH the hydride transfer occurs, which determines which hydrogen is transferred (Fisher *et al.* 1953).

[370] The identification of these residues as binding residues was partially locally done (based on physicochemical properties (charges and hydrogen bonds, in general) and consistency in distances) and partially from previous work, namely the papers associated with the DHFR structures in use. One residue that was used not only because of binding (from the backbone nitrogen) is glycine 117 in Amniote DHFR structures (position 245 in "Appendix K: Partial DHFR alignment", on page 384); it was also included because it is involved in positioning other residues via its tight turn with the immediately prior glycine (116 in Amniote DHFR structures). This GG pair is conserved in all DHFRs aligned (and apparently in at least some bacterial DHFRs as well, from a brief examination of the known structures), and in most examined structures has a conformation such that only glycines would fit (unusual phi/psi angles (Lovell *et al.* 2003), frequently combined with a cis peptide bond).

[371] This included residue numbers in the models; a file (e.g., "AO.NO.constraints.ascomycota.txt", available as a supplemental file) is thus needed for each set of models with a particular arrangement of gaps.

could "choose" whether to include NADPH restraints, and, if they were included, whether to use strict or non-strict restraints.)

2. The second set of restraints, between protein atoms, was more complex than the above (and, probably for this reason, more error-prone). The atoms restrained were pairs of alpha carbons or pairs of beta carbons (except that an alpha carbon from a glycine could pair with a beta carbon for another residue). The ranges[372] of distances for all such pairs were first found; these ranges were then tightened, by the program "find.distance.min.max.matrices.pl", using the techniques[373] of distance geometry (Havel 1998, 2007). The output from this program was then processed by "find.distance.deviations.2.pl" (run for each set of models with a given pattern of gaps), which looked for distances much above or below that expected for a particular length along a sequence[374] and tight ranges of distances around gap areas (the corresponding residues were

[372] This included the smallest distance seen, the largest distance seen, the smallest species median distance seen, the largest species median distance seen, and two compromises (as if doing a "non-strict" range) between these.

[373] Distance geometry, in the relatively simple form used here, looks at three points at a time. The range of possible distances between point A and point B, and that between point B and point C, should be already known; if so, then distance geometry enables the determination of the possible range of distances between point A and point C (the "triangle inequality"), or possibly the tightening of said range if it is already known. The present work took advantage of not only the distances gathered from the DHFR files, but:

- a reasonable range for alpha carbon-beta carbon distances (from `reduce` and the GROMACS force fields) of 1.526-1.54 Ang
- the VdW radii of carbons from `reduce; and`
- a (high) value for the maximum distance from one alpha carbon to another (1.53+1.33+1.47 Ang.), taken from the GROMACS force fields in use in this project.

[374] These were from formulae found by earlier work using the programs "check.main.chain.distances.2.pl" and "check.side.chain.distances.1.pl" on known DHFR and DHFR/TS structures along with a nonlinear least-squares equation solver. It would be preferable to use other protein structures as well for this, but the volume of data was already problematic for the equation solver program.

input from "find.pdbatom.stockholm.alignment.pl"). The result of this was a file used by "create.restraints.1.pl" along with, as before, a topology file.

It was unfortunately determined quite late, with the fungi/metazoa models, that the distance restraints of the second type were significantly too tight. Initially, this was thought to be a problem with one area being restrained, with a loop insertion in the DHFR of (some) invertebrates and fungi (in the middle of the sixth section of the alignment in Stockholm format - see "5. Alignment of central sequences", on page 336). This was deleted for the "full.partial" and "full2.partial" runs for fungi/metazoa, with resultant improvements in MolProbity scores (see "Appendix E: MolProbity results", on page 371). Upon seeing the Urascomycota MolProbity results, an attempt was made to do minimizations without the DHFR-only restraints. This attempt also yielded an improvement in MolProbity scores, but not enough for the Urascomycota results to be considered satisfactory, possibly due to earlier restraints having been too tight and simulated annealing (which might have loosened the structures) having initially failed (see "Simulated annealing when needed", on page 183). The question of how to set restraints is thus still in flux - other usage of MolProbity results (e.g., as per previous research using GROMACS (Flohil, Vriend, & Berendsen 2002), which would suggest concentrating restraints on problematic areas only) for the creation of restraints is one possibility (see "7. Model building", on page 345).

Non-frozen vacuum minimization

For all model creation rounds after Uramniota[375], two stages of (non-frozen) vacuum minimization were done, with NADPH being included (see "NADPH insertion", on page 156). See "Partially frozen vacuum/dry minimization", on page 167, for most information on how vacuum energy minimization was done; the following (aside from no atoms being frozen) are the differences in the stages of (non-frozen) vacuum energy minimization:

1. The first stage used the mdp file "vacuum.mdp". This file calls for an energy minimization with non-strict restraints on both the NADPH and (if applicable - see "Creation of restraints", on page 170) the protein (DHFR). The minimization terminated when either the maximum force (on any one atom) dropped below 187 kJ/(mol*nm)[376] or roundoff limits (for movement of atoms or reductions in energy) were reached; normally, the first happened.
2. The second stage used the mdp file "vacuum2.mdp", with strict restraints on both the NADPH and (if applicable - see "Creation of restraints", on page 170) the protein (DHFR). The energy minimization terminated when either the maximum force went to below 10 kJ/(mol*nm) - this is the default

[375] Initially, 3 stages of vacuum minimization were done (the additional mdp file used was "vacuum3.mdp"). From other work (Summa & Levitt 2007), this appeared to be inadvisable with the GROMOS96 43b1 vacuum force field. However, the HM_0.1 field developed in that paper for *in vacuo* minimization unfortunately does not appear to be available online, nor was the prospect of rewriting the conversion mechanism (see "Translations to/from GROMACS, PDB formats", on page 165) for that force field appealing given time constraints and potential errors. (The latter was particularly true with regard to NADPH nomenclature; indeed, we do not know whether the HM_0.1 force field includes NADPH.) The same was true of using the OPLS/AA (Das & Meirovich 2001; Jorgensen & Tirado-Rives 1988) force field, evaluated in that paper as second best. Therefore, minimizations done after the location of that research (all after Uramniota) involved only two rounds of (non-frozen) vacuum minimization.

value for GROMACS - or roundoff limits were reached; which happened first was variable.

Following the above, the protein was translated back from GROMACS to PDB format, run through reduce3, then put back into GROMACS format (see "Translations to/from GROMACS, PDB formats", on page 165) with the force field used being a modified[377] version of GROMOS96's 53a6[378] force field (Oostenbrink *et al.* 2004; Oostenbrink *et al.* 2005). The box[379] was created, with enough room to contain the water to be added (see "Addition of water", on page 178) - the size of 8.0 nm x 8.0 nm x 9.6 nm[380] was used in all but the first one or two (trial) modeling rounds - and the position of the protein and NADPH centered via GROMACS' "editconf" program.

---

[376] 187 is approximately (due to gaps varying between some models) equal to the number of residues plus 1 for the NADPH.

[377] As with the 43b1 force field, the modifications were to add some hydrogens to the NADPH that were either not added by `reduce` or did not translate properly to GROMACS; see patchfile "ffG53a6.hdb.patch" for the modifications.

[378] As with the 43b1 force field, the 53a6 force field explicitly includes all hydrogens except those attached to aliphatic (non-aromatic) carbons.

[379] The "box" is the computational mechanism for making the simulation more realistic by surrounding the protein (or other molecules) with, instead of empty space, copies of the molecules inside (e.g., the protein, solvent, and any ligands) - this setup is otherwise known as "periodic boundary conditions". (Using only a subset of the molecules in the “box” to be duplicated - e.g., solvent only - would be more complex and would leave a large hole in the water in the other boxes.) The box is the amount of room allowed for the molecules inside before the copies are encountered; molecules (e.g., water) that drift to the edges of the box are shifted to the directly opposite position. Please see the GROMACS manual (via http://www.gromacs.org), chapter 3 for more information on this, including pictures.

[380] The numbers used, besides considerations of the minimum size necessary, were selected so that, with a "fourierspacing" of 0.1 nm, the sizes of the FFT used for the Particle Mesh Ewald (PME) electrostatic (Coulomb) simulation were optimal for performance (being entirely products of the small primes 2, 3, and 5), as per FFTW's (Frigo & Johnson 2005) and GROMACS’ documentation.

Addition of ions if necessary

In some cases (depending on the sequence), it was necessary to add ions to reach a state of neutrality, both for reasons of biological realism (ions would be attracted to a charged protein/NADPH complex) and to prevent ionic repulsion problems with the minimization. Except for the first stages, in which Ca^{2+} was used[381], Na^{+} was used (or, for negative ions, which were occasionally necessary, Cl^{-}).

As per the procedure suggested for GROMACS' "genion" program, some water was first added; see "Addition of water", on page 178, although the "--maxsol" option was not used and the maximum distance from the solute was 0.9 nm (9 Ang.), due to 0.9 nm being the recommended minimum value for GROMACS' rcoulomb parameter. One or more of the water molecules (the O or H of said water molecule(s), to be precise, depending on the charge of the desired ion) was then selected for replacement by an ion or ions. Unfortunately, as noted in genion's documentation (which recommends random placement instead) and confirmed via experimentation, genion has problems determining the proper place to put ions based on electrical potential; the genion program was therefore not used.

[381] Calcium was used due to its presence in the chicken DHFR structure, 8DFR0. However, the calcium ion was rather far from the structure, and highly unlikely to interact with the structure; it appears to be a crystallization artifact. Therefore, the more common sodium was used after the recognition of this (when a model was encountered that needed an odd number of positive charges for neutralization). (Potassium would be another possibility, of course, and perhaps a biologically more reasonable one for an intracellular enzyme such as DHFR; however, the difference appears unlikely to matter.)

The structures including waters were instead converted back into PDB format for processing. The water atom(s) to be replaced were then selected, initially manually (based on closeness to the opposite-charged groups (either in the NADPH or in the protein), with distances determined by "check.atom.distances4.pl" or "check.atom.distances5.pl"). Later[382], the program "check.water.for.pos.ions.pl" was used, which decided upon the best location(s) for positive charges (negative ions were not needed for the latter stages) by:

1. adding up (the absolute value of) the (formal) negative charges divided by the square of the distance (as per Coulomb's law) to the possible location, then
2. subtracting the (formal) positive charges already present (including any added in an earlier iteration of the program) divided by the square of the distance to the possible location, and
3. using the location with the highest potential from the above for the placement of the ion.

In this procedure, the ionization state of the histidines had been determined by `reduce` earlier, using hydrogen bonding and steric realism. In all cases thus far, the histidines were uncharged; even if `reduce` does not take into account a molecule being actually in water, the addition of water is not anticipated to change this, since all questionable histidines examined were buried. Following

[382] The replacement with an automated process was for several reasons:
- Increasing consistency;
- Allowing for the electrical influence of already-inserted ions in where to put new ions;
- As the start of a replacement for the GROMACS genion program's nonfunctional code for taking into account electrical potentials;
- Reduction of manual labor (especially given its consequent increased chances of error).

the replacement of some of the water atoms with the desired ion(s), the remainder of the waters were removed, and the structure was converted back into GROMACS format (using the same force field as before - see "Non-frozen vacuum minimization", on page 174) for the addition of water (see below).

Addition of water

Proteins do not exist in a vacuum, but instead are surrounded by water[383] (Karlin, Zhu, & Baud 1999). It is thus advisable to surround them with water for energy minimization that is intended to do anything other than relieving steric hindrances and similar functions (Creamer, Srinivasan, & Rose 1995; Goldman, Thorne, & Jones 1998; Olivella *et al.* 2002; Tsai *et al.* 1999; Wako & Blundell 1994a). This process includes (given the considerable electrical impact of water) any Coulomb (involving charge, H-bonds, *etc.*) interactions (Eswar & Ramakrishnan 2000; Eswar, Ramakrishnan, & Srinivasan 2003; Farnum *et al.* 1991; Jones, B E *et al.* 1994). The addition of water was done by GROMACS' "genbox", with some modifications to the source code[384] and to the file (used to avoid steric clashes) "vdwradii.dat"[385]. The "SPC" (Berendsen *et al.* 1981; van der Spoel 2002a,

---

[383] This is true of the proteins we are examining, which are, for instance, not membrane proteins (see "1. Determination of central protein", on page 19).

[384] These modifications were to the source code file "addconf.c" in GROMACS' "src/tools" directory; please see the patchfile "addconf.c.patch". The modifications in question were to improve the performance of the "--maxsol" option, by deleting waters over the desired number from that option based on their closeness to the "box" and not, as previously, on essentially a random basis. It would admittedly be preferable to instead delete waters based on their distance from the solute (protein plus NADPH plus any ions, in our case), but:

- this would be significantly more complex to implement (particularly in a programming language, namely C, instead of the Perl used for most programming in this research); and
- the procedure used should give approximately the same result, provided the solute has been centered in the box, as was done for this research.

[385] The modifications to "vdwradii.dat" (in the GROMACS "top" directory, also the location of the force field files modified) were to make the radii used more realistic in terms of steric hindrance

2002b) model of water and GROMACS' "spc216.gro" (starting) model of water configuration were used. The amount of water added, the "--shell" option of 1.48 nm (14.8 Ang.) used, and the box dimensions used (see "Non-frozen vacuum minimization", on page 174) were chosen to try to have at least 2 layers of water on each side of the protein plus NADPH (Gerstein & Lynden-Bell 1993; Kahn 2007a). The value used for the "--maxsol" option (maximum number of water molecules added) was equal to the volume of the box minus the approximate[386] volume of the solute before water addition, divided by[387] 0.5 nm x 0.5 nm x 0.5 nm, the usual number of molecules added being something between 4000 and 5000.[388]

Minimization of water and other non-protein atoms

The water (and other non-protein atoms - the NADPH and any ions added (see "Addition of ions if necessary", on page 176)) was then minimized, with Coulomb

---

and to add radii for the ions in use; please see patchfile "vdwradii.dat.patch" for the modifications. The values added/changed were taken from the `reduce` (Word & Richardson 2006) source code (using the "explRad" value).

[386] The approximation used was to multiply each of the dimensions of the solute, as given by "editconf", together, then divide by 2 to account for that the solute's real geometry will be closer to that of a sphere than to a (fully-space-filling) box.

[387] This is the approximate minimum volume per water molecule in a realistic simulation (Kahn 2007a).

[388] Please note that the removal of the water - after all minimizations (and simulated annealing) were complete - was desirable for speed and reduction of output file size before analysis and (if needed) visualization of the results. Most analysis programs (e.g., MolProbity - see "MolProbity", on page 186) had difficulty analyzing the system with water. Moreover, all locally tried (Kahn 2007d) visualization programs - except for those in GROMACS itself - were not able to visualize the system with water at all (producing crashes). In contrast, GROMACS did not have any problems with minimizing with this amount of water in (very) reasonable amounts of time, although simulated annealing (see "Simulated annealing when needed", on page 183) was more time-consuming. (In other words, the title of one of the major GROMACS papers, “GROMACS: Fast, Flexible, and Free” (van der Spoel *et al.* 2005), is truthful.)

(normal electrical) interactions enabled, so that the configuration of the solvent[389] was more reasonable.

This minimization took place (for most models after the initial two or so trial rounds) with two stages:

1. The first stage used the mdp file "init.water.min.mdp". This file calls for an energy minimization:
    a. with the protein atoms "frozen" (prevented from moving, in all three dimensions), but other atoms not frozen;
    b. using the "steepest descents" minimizer (as earlier noted, a slow but very reliable minimizer, the latter characteristic being why it was chosen); and
    c. with restraints (non-strict; see "Creation of restraints", on page 170) on the NADPH.

    The minimization terminated when either the maximum force (on any one atom) dropped below 500 kJ/(mol*nm) or roundoff limits (for movement of atoms or reductions in energy) were reached; normally, the first happened.
2. The second stage used the mdp file "init2.water.min.mdp"; this file calls for the same energy minimization as the above, except:
    a. with all non-hydrogen (heavy) protein atoms "frozen", but other atoms not frozen (to allow for H-bonding and other interactions

[389] This minimization is also done with the other non-protein atoms unfrozen, of which:
- the NADPH had been previously minimized sans Coulomb interactions; and
- the ions had not been previously minimized.

between the hydrogen atoms (which were mostly added by `reduce`) and water, *etc.*); and

b. using the "conjugate gradients" minimizer (which, again as earlier noted, is faster but sometimes less reliable when far away from the energy minimum).

The minimization terminated when either the maximum force dropped below 56 kJ/(mol*nm) [390] or roundoff limits were reached; which happened first was variable.

As with all subsequent with-"box" minimizations, the maximum cutoff distances[391] used were the maximum usable with the size of box without the individual atoms[392] potentially interacting with their duplicates in the neighboring boxes. For more details, please see the ".mdp" files.

Full energy minimization

Prior to "full" (with water) energy minimization, some waters that appeared to be too far from the non-solvent molecules were frozen, to decrease the complexity of the problem[393]. Which water atoms/molecules were far away was determined

[390] The number 56 was chosen as the cutoff because that is the (approximate) number of atoms in NADPH in the force field (topology) in use.
[391] Above these distances, forces (VdW and Coulomb) are not used in detail.
[392] Of course, different (non-corresponding) parts of the molecules in each (duplicate) box may well have interacted.
[393] It is admittedly doubtful how much this did in some cases, in which few water molecules were frozen (less than 10 frozen for all atoms, although more with only the oxygens frozen). However, GROMACS did not have any time problems with the energy minimizations even without many waters frozen; see footnote 388, on page 179, for more commentary on this.

by the program "create.freezegrps.2.pl"; the criteria were that atoms at least 10 Ang. away from the closest solute molecule, and either[394]:

- significantly further away from any solute molecule than that; or
- very close to the box

would be frozen. This program also split the (water) atoms frozen into two groups, those:

- for which the entire water molecule was frozen (interactions between these were ignored for most purposes by GROMACS); and
- for which only part of the water molecule (the oxygen, so that the hydrogens could rotate for H-bonding optimization) was frozen.

Molecules other than (some) waters were not frozen for “full” energy minimization.

In the later rounds of model creation (see footnote 395, on page 181), full energy minimization used three stages:

1. Using the mdp file "init.full.steep.min.mdp"[395], energy minimization was:
    a. using the "steepest descents" minimizer (as earlier noted, a slow but very reliable minimizer, the latter characteristic being why it was chosen);
    b. with non-strict restraints on both the NADPH and (if applicable - see "Creation of restraints", on page 170) the protein (DHFR);

[394] The exact (numerical, *etc.*) specifications of these criteria differed somewhat between oxygens and hydrogens in water, to encourage freezing the oxygens in place while leaving the hydrogens free to rotate around to form a better H-bonding network.
[395] The first stage was not done on some earlier rounds, but appeared necessary from problems encountered with conjugate gradient minimization being unable to find an overly far away

The energy minimization terminated when either the maximum force went to below 10 kJ/(mol*nm) - this is the default value for GROMACS - or roundoff limits were reached (the latter being more frequent).

2. Using the mdp file "init.full.min.mdp", energy minimization was the same as with "init.full.steep.min.mdp", except for using the "conjugate gradients" minimizer (which, again as earlier noted, is faster but sometimes less reliable when far away from the energy minimum).
3. Using the mdp file "init2.full.min.mdp", energy minimization was the same as with "init.full.min.mdp", except with strict restraints on both the NADPH and (if applicable - see "Creation of restraints", on page 170) the protein (DHFR).

For all three, movement (of any type, including directional and angular) of the center of mass of the non-solvent molecules was cancelled out (thus allowing the choice of frozen water molecules to continue to be consistent with the goal of their being far enough away not to matter). For more details, please see the mdp files.

Simulated annealing when needed

Simulated annealing was performed at the fungi/metazoa ancestral sequence stage due to the negative trends seen in the MolProbity results (see under "8. Examination of models", on page 355). The initial try at simulated annealing unfortunately turned out to be done at too high a temperature to start (800K), due to using a (well-cited) paper as a source that used a temperature of 1200K

---

minimum, and was thereafter used for all "full" - with water - energy minimization.

(Berendsen *et al.* 1984)[396]. This error was determined[397] after the results from different simulated annealing runs were not alignable (within 2.08 Ang.[398] RMSD) to each other (despite identical sequences), nor to the original models.

The procedure used for simulated annealing for the more recent (and apparently more successful - see under "8. Examination of models", on page 355) simulated annealing runs was:

1. A previously energy-minimized (with strict restraints; see "Creation of restraints", on page 170) structure, with waters, was taken as the starting point.
2. "create.freezegrps.2.pl" (see "Full energy minimization", on page 181) was run, with the "$freeze_more" variable set, so as to cause it to freeze more of the waters, due to the increased amount of time needed for simulated annealing (as opposed to energy minimization with GROMACS). Interactions between fully frozen waters were ignored (to the degree possible).

[396] We note the year of the paper in question, and suspect that our normal practice of consulting, when possible, the most original paper on a subject was in error in this instance. That the paper in question was regarding a simulation of a much smaller peptide may also be a factor.
[397] Notable in hindsight was that a much larger proportion of the water was frozen by "create.freezegrps.2.pl" after simulated annealing (for energy minimization, due to being far from the solute(s); see "Full energy minimization", on page 181). It is fortunate that this was not a physical experiment, given the safety hazards of a steam explosion.
[398] This is the RMSD expected for 30% identity (Vogt, Etzold, & Argos 1995).

3. 3 parallel runs, with 3 different seeds (e.g., 173529) for randomization, were done, using "init.full.SA.min.[SEED].mdp" (e.g., "init.full.SA.min.173529.mdp"). This file called for the following run:
    a. Settings as to cutoffs and the presence of restraints (non-strict) were as per "Full energy minimization", on page 181;
    b. Simulated annealing temperatures were controlled via the Berendsen (Berendsen *et al.* 1984) method; separate "baths"[399] were used for (non-frozen) water and for the protein, with the first being much more tightly controlled;
    c. The annealing went from 298K (with starting velocities randomly determined, with the appropriate seed) to 373K (quickly) then, more slowly, to 0K;
    d. The annealing lasted 22000 steps (30.3 ps), of which the last 2000 steps would be at 0K (as a preliminary to energy minimization);
    e. Distance restraints (see "Creation of restraints", on page 170) were time-averaged[400], were set on "equal" weighting,[401] and were "mixed"[402].
4. A full energy minimization was then performed, with NADPH (to DHFR; see "Creation of restraints", on page 170) restraints only.

[399] The Berendsen method of temperature control effectively simulates the molecules being in a bath that is itself temperature-controlled. The molecules are not directly temperature-controlled, to prevent disturbing them too much.
[400] In other words, instead of the restraints being figured instantaneously, only if the *average* distance over a period (of simulation time) was outside of them was a force applied.
[401] Normally, distance restraints from one atom to multiple other atoms are concentrated, essentially, on the closest of the atoms (it assumes restraints are being used for NMR, in which case the needed force would decline as the 6^{th} power of the distance). Equal weighting means that, instead, each of the other atoms feels the same force from the restraint.

5. Currently, going further is waiting on better programs for handling averaging (which will be done with the previous level (Urdeuterostomia)'s structures as well as between those from all the usable Fungi/Metazoa runs); see "Assignment of initial coordinates", on page 150, and "7. Model building", on page 345. Following these steps and perhaps another energy minimization, another simulated annealing (with "strict" restraints) is planned.

Model building and sequence uncertainty

In most cases, as noted in "6. Determination of ancestral sequences" (on page 133), multiple possible ancestral sequences were determined. As well as with the model evaluation stage (see "MolProbity" below), some of these were also eliminated through the process of model building; if a model could not reasonably be built of a sequence, then it was considered an unlikely sequence. Please see "7. Model building", on page 345, for results and further discussion.

8. Examination of models

MolProbity

Model evaluation was primarily performed via MolProbity (Davis *et al.* 2007; Lovell *et al.* 2003), including the optional evaluation of bond distances and angles. This program (or, rather, complex of programs) was accessed via its webpage (http://molprobity.biochem.duke.edu) instead of downloading it, due to

[402] In "mixed" distance restraints, the restraint is only applied if *both* the current *and* the time-

the complexity of setting up webserver-based programs scripted using a language (PHP) not otherwise used in our laboratory. However, the addition of hydrogens via `reduce`[403] was performed locally. The evaluation from MolProbity included two categories of information about potential problems:

1. Findings of areas in which the model was physically strained; these indications included overly close atoms ("clashes", i.e., overlaps of VdW radii), beta carbon deviations (Lovell *et al.* 2003), bond lengths, and bond angles. These, if severe, were considered likely to indicate local minima in energy minimization, overly tight distance restraints, or (if in an area of uncertain sequence) a possible indication of what sequence was most likely. (For instance, clashes would tend to indicate that a residue was too large. Bond length/angle problems may indicate that, for instance, a proline should *not* be present (due to its lack of conformational flexibility) or a glycine *should* be present (due to its greater conformational flexibility). beta carbon deviations may indicate either of these (e.g., a sidechain that was too large would tend to push on the beta carbon's location, distorting it).)
2. Findings of areas in which the model, while not physically strained, was different from the geometry found in native structures, via examination of backbone (Ramachandran) and rotameric (sidechain) angles. (These were done using comparisons with a database, top500 (Richardson, D C & Richardson 2001), that does not include DHFR.) These findings were

---

averaged distances are outside of the range of the restraint.

[403] `Reduce` was run with a dot density of 100 per square Angstrom as an improvement on the default 16 per square Angstrom.

considered to be[404] indications of energy minimization artifacts (since said geometric aspects are not targeted for optimization by energy minimization), as well as the other possibilities suggested for category 1 (with backbone angle problems being analogous to bond length/angle problems and rotamer problems being analogous to clashes and carbon-beta deviations).

As well as the uses noted above, MolProbity's results helped determine:

- which models were used as templates for the next stage of homology modeling - this evaluation emphasized category 2 above, since these are not based on qualities directly targeted by energy minimization; and
- for a given residue, what weight was given to each model used, including mainchain and sidechain as separate categories of problems[405].

MolProbity analysis was done whenever this appeared likely to be useful, including in all cases before using a model as a template for further work, or even considering a model for such usage.

Residue volumes

Evaluation of residue volumes was considered, but not performed. Problems were noted with all the examined programs for volume checking:

- AtVol (Word 1999); while this program comes closest to matching our availability requirements, it uses explicit hydrogens; all currently published

[404] It is, of course, possible that the geometry seen is possible, but simply very rarely seen in the database in question. However, from the available results (see "Appendix E: MolProbity results", on page 371, for a summary), the number of these for the models was too high for such a missing data problem to be a reasonable explanation for *all* of the MolProbity findings.

[405] Problems with prolines or glycines were considered to be problems with both mainchain and

standard residue volumes were derived using implicit hydrogens, and are thus not comparable[406].

- calc-volume (Gerstein & Richards 2001; Tsai *et al.* 1999); this program appears to have a problem with overflows[407]. Since it is not open-source, no alterations on this program were made and it was not considered usable.
- VOLUME (Richards 1974); this program is stated on its distribution webpage (Biology 2006) to be out of date, with other programs[408] being preferable. The process of using it is also complex (involving running another program, ACCESS (Lee, B-K & Richards 1971), as well as removing hydrogens) and difficult to automate (it is meant for manual usage (with menus), not command-line, with the same being true for ACCESS).

Upon further consideration, it was concluded that volumes were not needed:

1. MolProbity's checking for "clashes" effectively checks for overly small residue volumes; moreover, it is capable of doing so for residues that are not 90% or more buried, unlike volume methods suitable for possibly-inaccurate structures (Gerstein, Tsai, & Levitt 1995; Kahn 2007f).
2. Voids, detectable via volume checking (looking for unusually large volumes of buried residues), may well differ between *C. albicans*' and other DHFRs[409]. In other words, the tightness of packing may differ between

---

sidechain atoms.

[406] Further research using atvol to derive standard volumes - e.g., from the top500 database - with explicit hydrogens may be indicated, particularly if atvol is clearly made available under GPL or equivalent terms.

[407] A "trap" is generated upon execution if it is compiled using gcc (Foundation 2002) with the "-ftrapv" option.

[408] The other programs referenced were found to be inadequately available.

[409] Due to the recency of the decision to use *P. carinii* as an intermediary target and the time pressures motivating that decision, adequate research to know whether this is likewise true of its

these DHFRs, since *C. albicans*' DHFR reacts differently to urea (which tends to unfold proteins), methotrexate is not a tight-binding inhibitor for it, and other differences (Baccanari *et al.* 1989; Duffy *et al.* 1987), and thus voids may differ. While these potential differences would make examination of alterations in voids of interest for *C. albicans* and perhaps its ancestors, they also would make differences in voids inapplicable as a criterion for model quality.

Solvent-exposed surface areas were examined using `probe`[410] (Word *et al.* 2000), comparing these to simulated fully-extended residues (Kahn 2006); model values for these generally corresponded well with existing DHFRs in (aligned) locations in which said DHFRs conserve the degree of solvent exposure - please see the “SA” and “S2” lines in the sequence alignment ("5. Alignment of central sequences", on page 336) for further information.

DHFR has not yet been conducted.

[410] `probe` was run with a setting of 100 dots per square Ang. instead of the 16 dots per square Ang. default, to improve on accuracy. This was particularly of interest for distinguishing essentially completely buried residues (those without even 1 “dot” placed by `probe`); these are distinguished in the “S2” lines of the alignment (see above) by a “B”.

Chapter 4: Results, Discussion, and Future Work

Note that with regard to "Future Work", a number of other places in this dissertation (particularly footnotes[411]) discuss improvements that could possibly be made.

1. Determination of central protein

The usage of DHFR for the central protein appears to be a reasonable choice for the criteria of this research, although less-difficult proteins (e.g., ADH1) may be of interest for some further research, such as that regarding gap determination (see “Discussion and future work”, on page 344).

2. Determine sources for phylogenetic sequences

The database of structures versus species is available as supplemental files[412] "swissprot.scop.species2.txt" and "known.species.txt" (the latter being manual additions), and may be the subject of a future publication comparing its results with those from other sources. Where such comparisons indicate that another database[413] contains incorrect information about structures and species, we will attempt to notify the curators of the other database.

---

[411] See, for instance, footnotes 34 (on page 20), 116 (on page 62), 230 (on page 111), 234 (on page 113), 277 (on page 130), 311 (on page 145), 326 (on page 153), and 334 (on page 156).
[412] They are also available under http://cesario.rutgers.edu/easmith/research/species/.
[413] These databases will include only ones that are current, openly available, and reproducible under terms like ours (see “Choice and Availability of Programs and Data”, on page 43) or with fewer constraints (e.g., public domain, as for U.S. government publications).

3a. Creation of a rough starting tree

The original starting tree is unfortunately too large for inclusion[414] (please see Figure 4.1, on page 193, for an example of why; note that this is only showing the fungi in the starting tree). However, subsets of it (as the tree 1 arrangement) will be shown under "First round of tree rearrangements", on page 203. The technique of using quartet[415] testing versus a “trusted”[416] but incomplete tree appears promising, and may be a new idea. However, it is concluded that it would be better to do any "blurring" desired on a manual (or at least semi-manual) basis[417], unless one had more than one possibly trusted tree[418] with which to do a more well tested consensus algorithm than the current quartet implementation[419].

---

[414] A version of it in PHYLIP format is available in the supplemental file "trees.tar" (in UNIX “tar” format), named “original.round1.phy”; it is also available via http://cesario.rutgers.edu/easmith/research/trees/original.round1.phy.

[415] Other varieties of testing for congruence may also be helpful; any that, like quartets, are suitable for assisting with supertree construction (e.g., quartets from a trusted tree can be combined with those from a new tree) may be preferable.

[416] By a “trusted” tree is meant one that is unlikely to have significant incorrect classifications - in particular, is unlikely to have classifications that would be considered unreasonable by those knowledgeable in the field.

[417] On the other hand, at least with regard to completely manual “blurring”, the difficulties seen in the present work with manual tree rearrangements (see “Future work”, on page 334) should be kept in mind.

[418] In other words, more than one tree showing reasonably supported evolutionary hypotheses.

[419] On the other hand, it is possible that the erroneous quartets were introduced by the “cleanup” process - see under “Usage of quartets”, on page 75. A more restricted version of this process, or avoiding it altogether, may cure the problems seen.

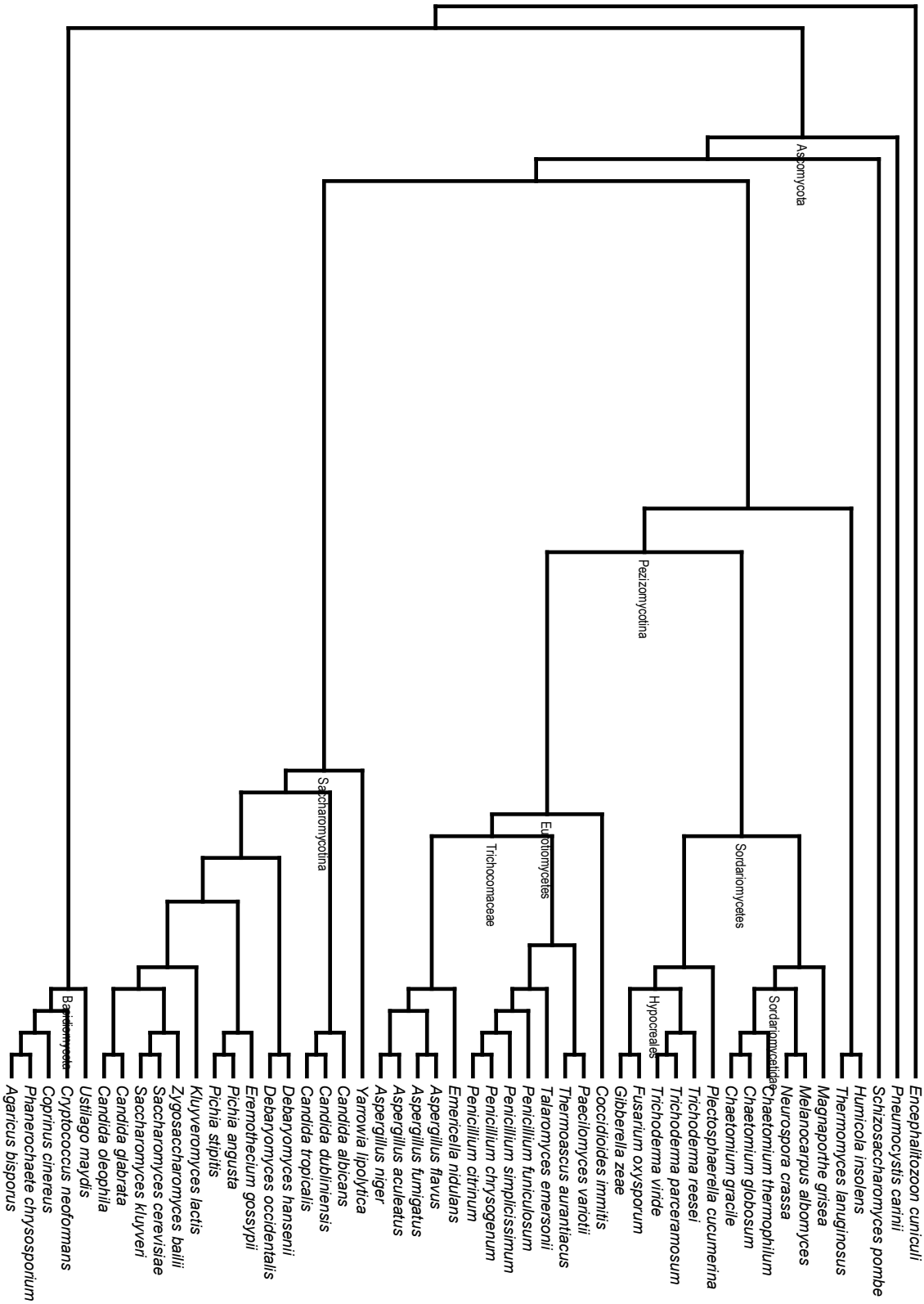


Figure 4.1: Fungi-only subset of starting tree (without distances)

3b. Alignment of other sequences

The database of alignments can be found at http://cesario.rutgers.edu/easmith/research/alignments/; a backup can be found in the supplemental file "alignments.tar", in UNIX "tar" format. In this page, "uncertain" areas are in black, while "nonstruct" areas are in red; reliably aligned areas are in blue. This database does not include DHFR/TS alignments; this is due to both the size of the alignment and its problematic nature in some areas (see “5. Alignment of central sequences”, on page 336). For future work in this area, an examination of HOMSTRAD for areas of uncertain reliability (see "Evaluation of structural alignment reliability", on page 84) would perhaps be helpful, although it is possible that the comparisons of it with other structural alignment methods will provide an adequate check in many cases. Improvements in the comprehensibility of the programs involved would be valuable. Some of the changes discussed below under "5. Alignment of central sequences" (on page 336) may be helpful for improvements of this database, especially with regard to alignments in the non-"struct" areas. Also in mind for improvements of this database are displays of the structural superimposition of the (3D) structures aligned, particularly those aligned locally, perhaps with “struct” and “uncertain” areas indicated.

4. Tree refinement

Following further testing (see below) and/or improvements when applicable, we will submit the changes to MrBayes to the MrBayes authors. With regard to the

covarion option (see footnote 200 under “MrBayes code alterations”, on page 99) and the below results, runs in which the use of covarion was attempted[420] were not used[421] if they gave either significant numbers[422] of covarion-related errors (e.g., LIKE_EPSILON[423]) or any large (possibly) covarion-related errors (e.g., positive log likelihoods).

Simulated Annealing (SA)

SA testing was done with 3 runs with and 3 runs without using SA, using 3 datasets[424] (the first was eukaryota, the second was bacteria, the third was archaea plus non-fungi/metazoa/plant eukaryota), without topology variations, with the same randomization seed per dataset. The results indicated acceptance ranges of "moves" were, at least for archaea and - in general - eukaryota,

[420] After the initial problems were noted with the covarion option, some runs were done while attempting to find ways to avoid these while still making use of it. These attempts involved:
- ultimately unsuccessful (except for the full usage of double precision and those noted in footnote 423 for “LIKE_EPSILON” errors) changes to MrBayes; and
- alterations to “combine.structural.align.groups.nexus.pl” in order to reduce the use of covarion for datasets that appeared to cause more problems with it (including ones with very low or high variability in particular “partitions” and ones with low amounts of data (either sequence length or number of sequences)).

[421] Fortunately, with the exception of one case of a positive log likelihood, such errors generally happened early enough that the run could be manually interrupted.

[422] For instance, if such errors were seen after the “burnin” phase (or, indeed, at any point after the first quarter or so of the run) or if there were more lines on a screen indicating said errors than those indicating a normal run (reporting the current log probabilities), the run was considered unreliable.

[423] “LIKE_EPSILON” errors originally resulted in significant inaccuracies in MrBayes, in that instead of the probabilities in question being treated as very small (close to 0, the cause of the error), they were treated as very large. This problem has been corrected in the local version, but runs generating significant numbers of LIKE_EPSILON errors have still not been considered valid, given the evident impact of roundoff or other errors.

[424] In other words, each dataset was processed twice - once with SA, once without SA. The starting conditions - aside from setting whether SA was used - were identical between the two runs for each dataset.

better[425]. Moreover, for 2 out of 3 cases, more branch length variances were above 0, indicating more branch lengths were tried[426]:

Subset	Number of branch lengths variable	
	Non-SA	SA
Eukaryota	3	**22**
Bacteria	**8**	5
Archaea	0	**3**

[425] By better is meant more in the desired ~10/20-70% range (Huelsenbeck *et al.* 2006; Ronquist 2005). For the detailed results, see supplemental files "new.SA.archaea.xls", "new.SA.bacteria.xls", and "new.SA.eukaryota.xls" (also available via http://cesario.rutgers.edu/easmith/trees/new.SA.archaea.xls, http://cesario.rutgers.edu/easmith/trees/new.SA.bacteria.xls, and http://cesario.rutgers.edu/easmith/trees/new.SA.eukaryota.xls). Please see "Appendix J: MrBayes review/explanation", on page 379, for more on why there is a desired range of acceptances.

[426] If a branch length has a 0 variance, then evidently no "moves" succeeded in trying to alter it; it appears doubtful that the ending lengths are completely correct (especially given that other parameters of the tree were being altered), so this is not a favorable sign. (It would admittedly be preferable also to have, for instance, some comparisons between the branch lengths ultimately derived and those from the final tree, to see whether SA improved the former relative to the latter. This would, however, best be done only after checking to see how reliable the distance determination methodology actually was; see footnote 232 under "Tree distances", on page 113.)

The log probabilities[427] were as follows (with a burnin[428] of 1500, total run 3000 samples - i.e., from 300,000 generations):

Kingdom	Mean Type	non-SA	SA
Eukaryota	Arithmetic	-205,190.73	**-173,579.22**
	Harmonic	-211,723.79	**-183,572.94**
Bacteria	Arithmetic	**-47,037.96**	-53,063.17
	Harmonic	**-49,644.47**	-56,796.73
Archaea	Arithmetic	**-24,176.70**	-24,395.27
	Harmonic	**-24,181.33**	-24,417.70

The CPU time[429] was increased by, at most, 200 seconds with SA, which cannot be described as a significant difference for a multi-day run.

While these results are unfortunately equivocal (the Bacterial subset is not improved, and the evidence on the Archaeal subset is equivocal[430], with a small difference in the wrong direction for log probabilities), we concluded that it was advisable to use SA with later stages, given that the Eukaryotal subset showed

[427] In **boldface** are the log probabilities indicating a higher likelihood as compared to others. This may be done for more than one, if the results are inconsistent or if more than one hypothesis' test is found in the table (e.g., for the variations on tree results, if two different, not inconsistent with each other variations on the original tree appeared to be correct). This formatting will be followed for subsequent tables, unless noted otherwise. Both the arithmetic mean and the harmonic mean are given; the arithmetic mean is more readily comprehensible and appears to be more stable, while the harmonic mean is preferable (Huelsenbeck *et al.* 2006) for the comparison of models. It is likely that the most reliable comparisons are when both means agree. One improvement on MrBayes' current reporting on the log probabilities would be giving standard deviations (although the applicability of these to (log) probabilities is somewhat unclear, and it is dubious whether the log probabilities would be normally distributed) or other numerical indications of variability (e.g., the 5th and 95th percentiles, provided there were sufficient samples available). Additionally helpful may be the output of median values. This is a matter for future work. (As well as for comparison purposes such as these, such information may be useful in deciding at what point a run has "stabilized", for purposes of determining the best "burnin" number to use - see footnote 428.)

[428] "Burnin" is used for both "sumt" (which extracts the trees generated during the course of a run and puts them together (including from multiple runs)) and sump (which extracts the mean probabilities for runs and the values found for various parameters). It is the number of samples from the start of a run that are skipped, to avoid collecting data from a period of time in which MrBayes' MCMC is likely to be still searching around for the solution and/or overly influenced by the starting conditions. Please see "Appendix J: MrBayes review/explanation", on page 379, for more information, including regarding "burnin" with "SA" and "Adapt".

[429] I.e., the computer runtime if no other programs were running.

[430] It is possible this is due to problems with the tree; please see footnote 470 under "Tree search

improvements, and this research focuses on Eukaryota. Some adjustments were made to the degree to which SA flattened out the probabilities, in addition to combining SA with sliding window/multiplier adaptations (see "Adaptation", on page 199); it appears that whether it is advisable to use SA and to what degree it is advisable to use it is unfortunately somewhat dataset-dependent. To be noted is that Eukaryota had, on the average, significantly more sequence data per species as compared to Archaea and Bacteria, so flattening out already-uncertain probabilities may be disadvantageous overall with these datasets. Moreover, the utility of SA may differ somewhat depending on what move's probabilities are being adjusted (with some interaction with whether sliding window/multiplier adaptation is being used; see "Adaptation", on page 199). It also appears likely that local minima are most problematic with tree topology variations[431], making SA usage with such (not done during this testing) most likely to be advantageous. Further research on this is suggested, including before any (methodological) publications concentrating on these changes to MrBayes. (For instance, one could check to see if the removal from the database of sequence data from some proteins harms SA-enabled MrBayes more than it does non-SA-enabled MrBayes.)

---

with Eukaryota (subset)", on page 303.

[431] Such variations are inherently non-continuous in nature (at least if not combined with tree distances), and affect a large number of other parameters.

Adaptation

In addition to SA, adaptation (Corana *et al.* 1987) of sliding windows[432] and multipliers[433] ("Adapt") was also tried for the parameters[434] for appropriate moves, because:

- SA's effects were primarily limited to the initial portions of runs and it was contemplated that it might be desirable to adjust the windows for the earlier portions of the runs, while adjusting them back to the starting values for later portions of the runs (but always stopping adjustments prior to the set burnin period's end), if this seemed desirable based on the acceptances;
- Of the general uncertainty in setting windows and multipliers via the "props" command (see footnote 204 under “MrBayes code alterations”, on page 101).

Time limits forced the testing of the initial (without multiplier adaptation) code versus non-adaptive code on one dataset only, namely one that included fungi/metazoa with known DHFR sequences plus a few other eukaryota[435]. The

[432] For moves involving a “sliding window”, an existing parameter is adjusted up or down by a randomly determined amount; the limits of this amount are determined by the sliding window size for that move. (Adaptation alters the sliding window size.) For instance, if the current parameter value was 0.2 and the sliding window size was +/- 0.1, then the move could try parameter values between 0.1 and 0.3. Please see “Adapt and SA”, on page 381, for more information.

[433] Adaptation of multipliers was done using log scaling to be able to use the sliding window algorithm, which assumes a linear relationship. (Moves using a “multiplier” are those that, instead of adding or subtracting a random number as per a “sliding window”, multiply or divide the current parameter by a random number (with its limits analogous to those for a sliding window size).) Again, please see “Adapt and SA”, on page 381, for more information.

[434] In this, the sliding window sizes and multiplier sizes are adjusted so that the percent acceptances of the moves are more within the recommended range, 10/20%-70% (Huelsenbeck *et al.* 2006; Ronquist 2005).

[435] The species were: *Anopheles gambiae, Apis mellifera, Arabidopsis thaliana, Bos taurus, C. briggsae, C. elegans, C. albicans, C. glabrata, Canis lupus, Coprinus cinereus, Cricetulus griseus, Cryptococcus neoformans, Cryptosporidium hominis, Danio rerio, Debaryomyces hansenii, Drosophila melanogaster, Drosophila pseudoobscura, Eremothecium gossypii, G. gallus, Gibberella zeae, Homo sapiens, Kluyveromyces lactis, Macaca mulatta, Mesocricetus auratus, Monodelphis domestica, Mus musculus, Neurospora crassa, Pan troglodytes, Pichia stipitis, P. falcip., P. carinii, Rattus norvegicus, S. cerevisiae, S. pombe, Strongylocentrotus purpuratus, Sus scrofa, Tetraodon nigroviridis, Tribolium castaneum, Ustilago maydis, Xenopus*

results were considered promising for variable branch lengths (in terms of having some variation in branch lengths, but not extreme levels of variation):

- non-Adapt: 5 variable branch lengths, but one of these had a variance of over[436] 1.5 and only 1 other had a variance of 0.01+;
- Adapt: 4 variable branch lengths, none with a variance over 0.6 and with 3 having a variance of 0.01+.

Acceptances (data not shown) were more difficult to evaluate, partially due to some moves being enabled for adaptation and some not. It appears likely that the Adapt-enabled move for "proportion invariant" was assisted, but the other moves (which in general were not Adapt-enabled) were sometimes harmed; this is made additionally difficult to interpret by that:

- subsequent code incorporated more moves being "Adapted";
- subsequent work did not use the covarion settings (see page 99, footnote 200), and the "switch" rates for covarion were among those that were apparently harmed by adaptation;
- subsequent coding altered the relationship between Adapt and SA[437]. This relationship may have caused problems prior to this revision for the SA-enabled but (at the time) non-Adapted moves (e.g., at the time the "Node Slider" move (adjusting branch lengths) was not Adapt-enabled and was apparently harmed by using SA and Adapt together).

---

laevis, Xenopus tropicalis, and *Yarrowia lipolytica.*

[436] Given that the square root of 1.5 - the standard deviation of the branch length - is over 1, and a branch length of 1 implies that every position changes, then a change by more than 1 in the branch length indicates some variety of problem.

[437] It is not desirable for both Adapt and SA to be affecting the probability of a move being accepted, given that SA's effects are diminishing while the Adapt effects are (until revised) lasting. The original coding inhibited the use of SA to a considerable degree when any moves

The log probability comparison, which is favorable for Adapt, is as follows (the burnin was 1500 with 2000 samples (200,000 generations)):

Mean Type	SA, non-Adapt	SA, Adapt
Arithmetic	-112,698.10	**-109,051.35**
Harmonic	-118,007.81	**-115,407.24**

Again, no significant difference was seen in CPU time consumed. It would again be greatly preferable to perform more testing on this prior to any (methodological) publications focused on these MrBayes code changes.

Tree results

The below results are arranged in the (approximate) order in which the findings were made. Note that some earlier tree search results are not shown; problems with these resulted in the conclusion that tree rearrangements were needed as a primary means of tree determination, although some results from them as to problematic species were used, as noted earlier[438]. Also note concerning the reporting of the number of amino acids that:

1. This measurement is of the total number of positions included in the alignment file as potentially at least *slightly* useful (see under “Species subsets”, footnote 211, on page 104); it is not a measure of, for instance, how many sequence positions were in common between species. The development of such a measure, perhaps based on informational entropy

---

were Adapted.

[438] These problems (frequent inconsistent placement of species between runs) were analyzed via “compare.trees.problems.pl”, as noted:

- under “Species, polymorphism reduction”, on page 71;
- in footnote 213 under “Species subsets”, on page 104;
- in footnote 476 under “Tree search with Mammalia (subset)”, on page 316.

(Lin 1991; Liu, X Z *et al.* 2003; Yona & Levitt 2002) *in common*, may be of interest.

2. This number was divided by 3, to account for the tripling of positions due to accommodations for ADH1's isozymes (see under "Other proteins used", on page 58); this admittedly undercounts the contribution of ADH1 slightly for primates with more than one such isozyme known.

In the below, the "subsets" are different collections of species (see "Species subsets", on page 101), with (consequently) different subsets of proteins used. In the tree pictures, "phylogram" indicates that distances are shown, whereas "cladogram" indicates that no distances are shown[439]. A tree noted as "X only" or "X only shown" (e.g., "Eukaryota only") is a subtree of a larger tree on which a tree rearrangement/search has been conducted[440]. The figure names are abbreviated descriptions of this information (e.g., 4.T.r1.s2.c.p means "Chapter 4,

[439] There are three reasons why branch lengths may not be shown:
- For the final tree, in order to assist in seeing the branching order even when the distances are small.
- For trees from tree rearrangements, because the transfer of branch length data from one tree (i.e., the final tree) onto another with a different topology tends to result in distortions in the branch lengths, such as some being very close to 0 (see item 7 under "distances", on). Branch lengths derived solely from runs with a single topology other than the final one (or a subset of the species in the final one) are likely to be unreliable compared to the branch lengths from the final tree, which has many more runs put together. This is particularly the case for tree rearrangement rounds, since the initial set of distances were arbitrary for these in order not to bias the result towards the topology from which the distances were derived.
- For trees in general, to make it easier to see data on inside/inner nodes (e.g., group labels like "Eukaryota" or validity information like "0.99").

[440] In other words, other species than Eukaryota were involved in a rearrangement/search presented in a "Eukaryota only" tree, even though only Eukaryota are shown in the particular tree picture. This procedure is for two reasons:
- For trees with branch lengths, to enable examination of the smaller branch lengths, such as those for within Mammalia on a tree with Archaea present;
- For trees without branch lengths, to enable easier comparisons between the results of tree rearrangements. (That this appears to be necessary for clear comprehension of the results is an additional argument in favor of further automating the rearrangement process; see "Future work", .on page 334. Note that the original work was done without such aids; this may be partially responsible for some of the errors made.)

Tree Figure, Rearrangement Round 1 Subset 2 of Current/Final Tree, Phylogram").

With regard to the tree display pictures, some problems may be seen:

- Inability to present (readable) group[441] information for inside[442] nodes; and/or
- Group information that obscures some species names.

We apologize for this, but the tree display programs that have been located have considerable limitations. For some trees, species names are not italicized, for readability reasons.

First round of tree rearrangements

The possible tree rearrangements (hypotheses about organismal descent) tested by each phylogenetic comparison done for round 1 are as follows:

- 1 versus 2|3|4 - This set compared two hypotheses as to the arrangement of Metazoa:
  - Coelomata, Pseudocoelomata[443], and Acoelomata (e.g., *Schistosoma*); this may be considered the "classical" arrangement (Bischoff *et al.* 2004; Jones, M & Blaxter 2005; Philippe, Lartillot, & Brinkmann 2005; Wheeler *et al.* 2000). In this, Coelomata are divided into Protostomia (e.g., Insecta) and Deuterostomia (e.g., Vertebrata). In this taxonomy, Acoelomata is generally thought of as branching off first, as the "simplest" organisms (Philippe, Lartillot, & Brinkmann 2005).

[441] For groups, see under "Appendix I: Species groupings used", on page 378.
[442] By "inside" is meant "not terminal", with "terminal" meaning associated with sequences (i.e., a

- Division of (non-Deuterostomia) Metazoa into Ecdysozoa and Lophotrochozoa (Jones, M & Blaxter 2005; Philippe, Lartillot, & Brinkmann 2005; Ruiz-Trillo *et al.* 1999; Telford, Wise, & Gowri-Shankar 2005), which divide between them Protostomia, Pseudocoelomata, and Acoelomata. In this, the resulting divisions are of Deuterostomia on one branch and Ecdysozoa (with Nematoda and Insecta) plus Lophotrochozoa (with Acoelomata and the Protostomia not grouped with Ecdysozoa) branching together on the other. This hypothesis was the one initially assumed[444] due to prior evidence (Philippe, Lartillot, & Brinkmann 2005; Ruiz-Trillo *et al.* 1999; Telford, Wise, & Gowri-Shankar 2005).

The alternatives to the second hypothesis ("Ecdysozoa and Lophotrochozoa") were variations on the branching order of the first arrangement:

- Tree 2: This variant had Pseudocoelomata and Acoelomata together.
- Tree 3: This, the most "classical" variant (Philippe, Lartillot, & Brinkmann 2005), has Acoelomata branching off first.
- Tree 4: This variant had Coelomata and Acoelomata together (which in some respects is the least "classical" of 2, 3, and 4).

Note that some - not all - subsets could not distinguish between 2, 3, and 4 (the tree for 3 and 4 was identical to that for 2 for these), even if they could distinguish between those (shown as 2|3|4) and tree 1.

---

species or outgroup (see "Further sequence processing: Group sequence creation", on page 96)).

[443] Pseudocoelomata include Nematoda such as *C. elegans*.

[444] One possible variation with regard to this arrangement is with the position of Mollusca, which tree 1 placed within Protostomia (as per the NCBI taxonomy); however, there is unfortunately only one member of Mollusca in the dataset, *Ommastrephes sloani*.

- 1 versus 5 versus 6 - This set compared positions for *Candida* species:
  - Tree 1 (original/starting): *C. albicans* and others known to have a differing CUG codon[445] (coding for Serine instead of Leucine (Sugita & Nakase 1999a)) than the "standard" one together, and other *Candida* species (*glabrata* and *oleophila*[446]) closer to *S. cerevisiae*
  - Tree 5: *C. glabrata* and *oleophila* with *C. albicans*
  - Tree 6: *C. albicans* (and others) with *C. glabrata* and *S. cerevisiae*
- 1 versus 12 versus 13 - This set compared whether *D. discoideum* and *E. histolytica* are:
  - Branching off together prior to the fungi/metazoa divergence, as with the original tree (1)
  - Closer to fungi than to metazoa (12)
  - Closer to metazoa than to fungi (13)
- 1 versus 15 - This tree rearrangement was of bacterial groupings (at the level of significant group sizes, e.g., Proteobacteria), comparing the initial arrangement[447] with that suggested by some prior research (Gupta 1998, 2000, 2001, 2005, 2007).

---

[445] It should be noted that:
- This change is, in this instance, unlikely to evolve more than once, due to being a more complicated form of codon change than most, since more than one base in the tRNA required changing (Massey *et al.* 2003)
- This grouping includes ones with variable codon usage, although this is admittedly argued against by some evidence regarding other genetic changes (Gibb *et al.* 2007) - however, it appears that the trait of variable codon usage may itself be, not a transitional state, but adaptive in and of itself, notably in *C. albicans* itself (Gomes *et al.* 2007).

[446] The latter (*oleophila*) was thought at the time, due to the information from the NCBI taxonomy, to have the standard genetic code; since then, research has been located (Sugita & Nakase 1999a) indicating it has the "*Candida*" genetic code as well. Note that many more sequences are known for *C. glabrata* than *oleophila*.

[447] The initial arrangement used was that of the processed (see "Appendix D: NCBI taxids and alternate species names", on page 370) NCBI taxonomy, with polytomies resolved using other

Initially, it had been intended for more hypotheses than the above (all in Eukaryota - the testing of 1 versus 15 was not supposed to be a major component of the research) to be tested; copying errors (see "Future work", on page 334, for more commentary) have prevented the analysis of these added hypotheses for this stage. (Some of the added hypotheses were tested later, under "Second round of tree rearrangements", on page 265).

Subset 2: Some Eukaryota, Bacteria

In the first round, subset 2 had 8433 amino acids, and 22 proteins (with ADH1 treated as 1). The runs used 200000 generations (2000 samples) with a burnin as shown below in the log probability table; the trees[448] are below the table (pages 208-220):

Phylogeny Tested	Burnin = 1000		Burnin = 1500	
	Arith. M.	Harmon. M.	Arith. M.	Harmon. M.
1 (original):	**-106,849.61**	**-121,181.63**	**-106,849.61**	**-108,012.49**
12:	-120,093.52	-147,699.16	-120,093.52	-120,459.95
13:	-128,762.34	-129,212.51	-128,762.26	-128,995.56
15:	-145,923.71	-163,651.84	-145,923.71	-146,345.18

research (see "Appendix C: Other sources for initial tree", on page 369).

[448] This subset's bacterial species are focusing on non-Proteobacteria, with a few representative Proteobacteria chosen by the subset species selection process (see "Species subsets", on page 101) as being close to the other species (and the root of the tree). Note that a number of species (primarily bacterial) were removed between this subset's creation and that of the final tree (for reasons ranging from the deletion of some proteins (see "Appendix F: Proteins removed", on page 373) to increased tightness of species selection (see "Species, polymorphism reduction", on page 70)). These were: *Agaricus bisporus, Bacillus agaradhaerens, Bacillus clausii, Bacillus megaterium, Cellulomonas fimi, Clostridium acetobutylicum, Clostridium stercorarium, Clostridium thermocellum, Micrococcus luteus, Nonomuraea flexuosa, Phanerochaete chrysosporium, Streptomyces avermitilis, Streptomyces olivaceoviridis, Streptomyces thermoviolaceus, Streptomyces viridosporus, Thermobifida fusca*, and *Xylanimicrobium pachnodae*. Of these (which were not included in the final tree), the following were converted into a "fungi" outgroup for display purposes for the "current" trees: *Agaricus bisporus, Coprinus cinereus*, and *Phanerochaete chrysosporium.*

Concerning which species or labeled species groups are rearranged (relative to tree 1) for each tree:

- 12 and 13: *D. discoideum* and *E. histolytica*
- 15: Firmicutes, Actinobacteria, Thermus-Deinococcus (*Thermus aquaticus*, *Deinococcus geothermalis*, *Deinococcus radiodurans*), Proteobacteria.

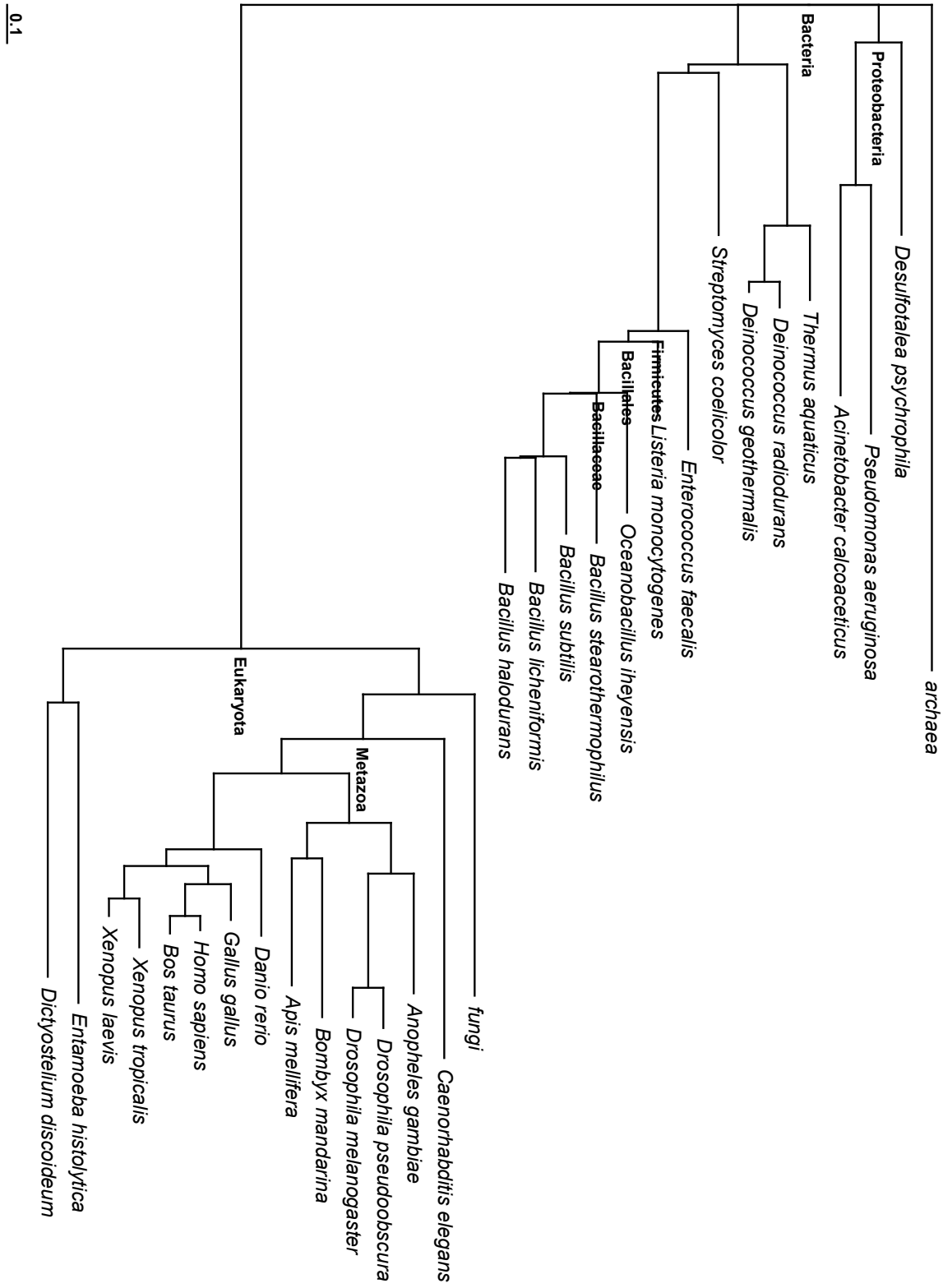


Figure 4.T.r1.s2.c.p: Round 1 subset 2 of final, phylogram

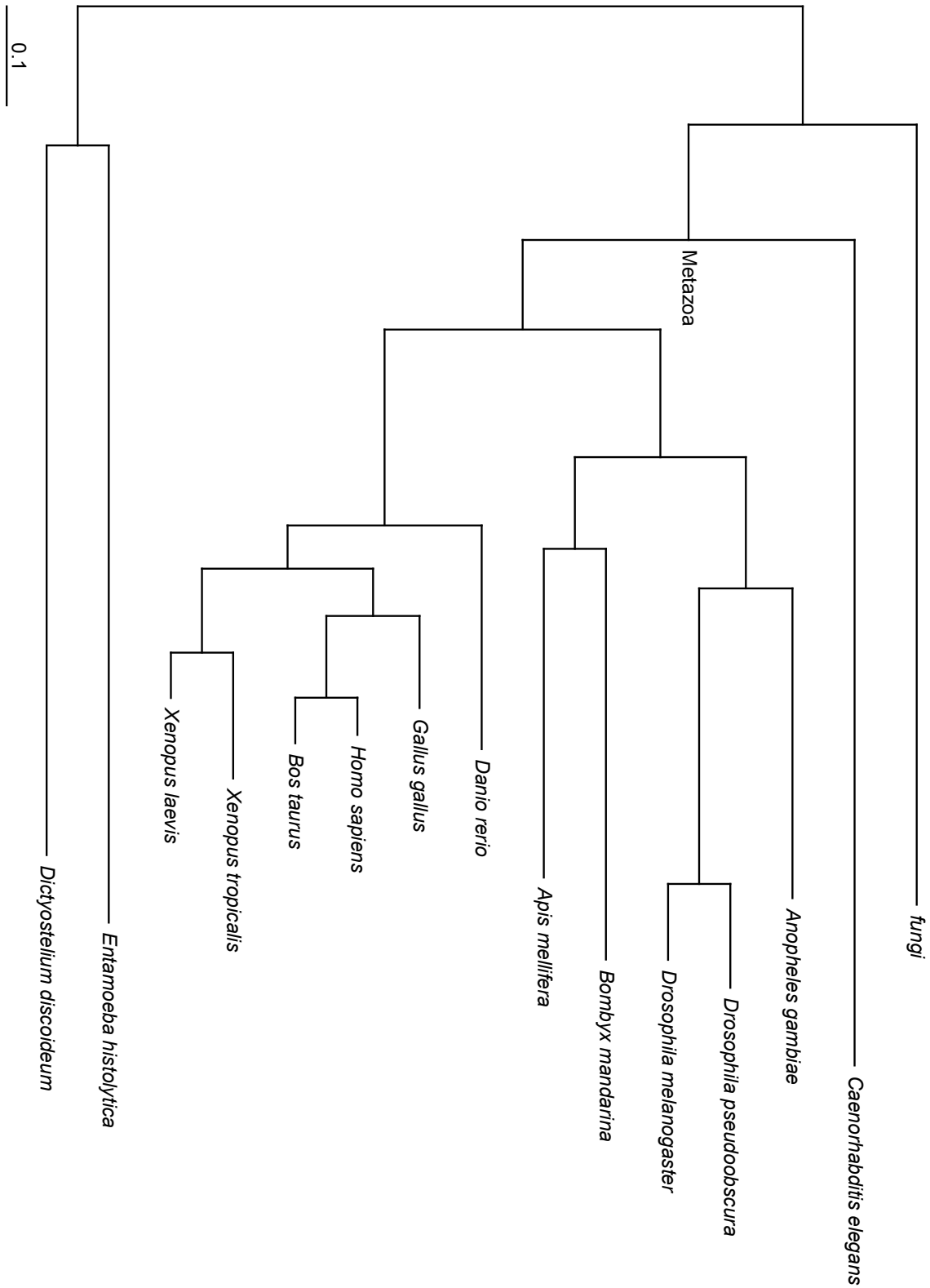


Figure 4.T.r1.s2.c.p.eukaryota: Round 1 subset 2 of final tree, Eukaryota only shown, phylogram

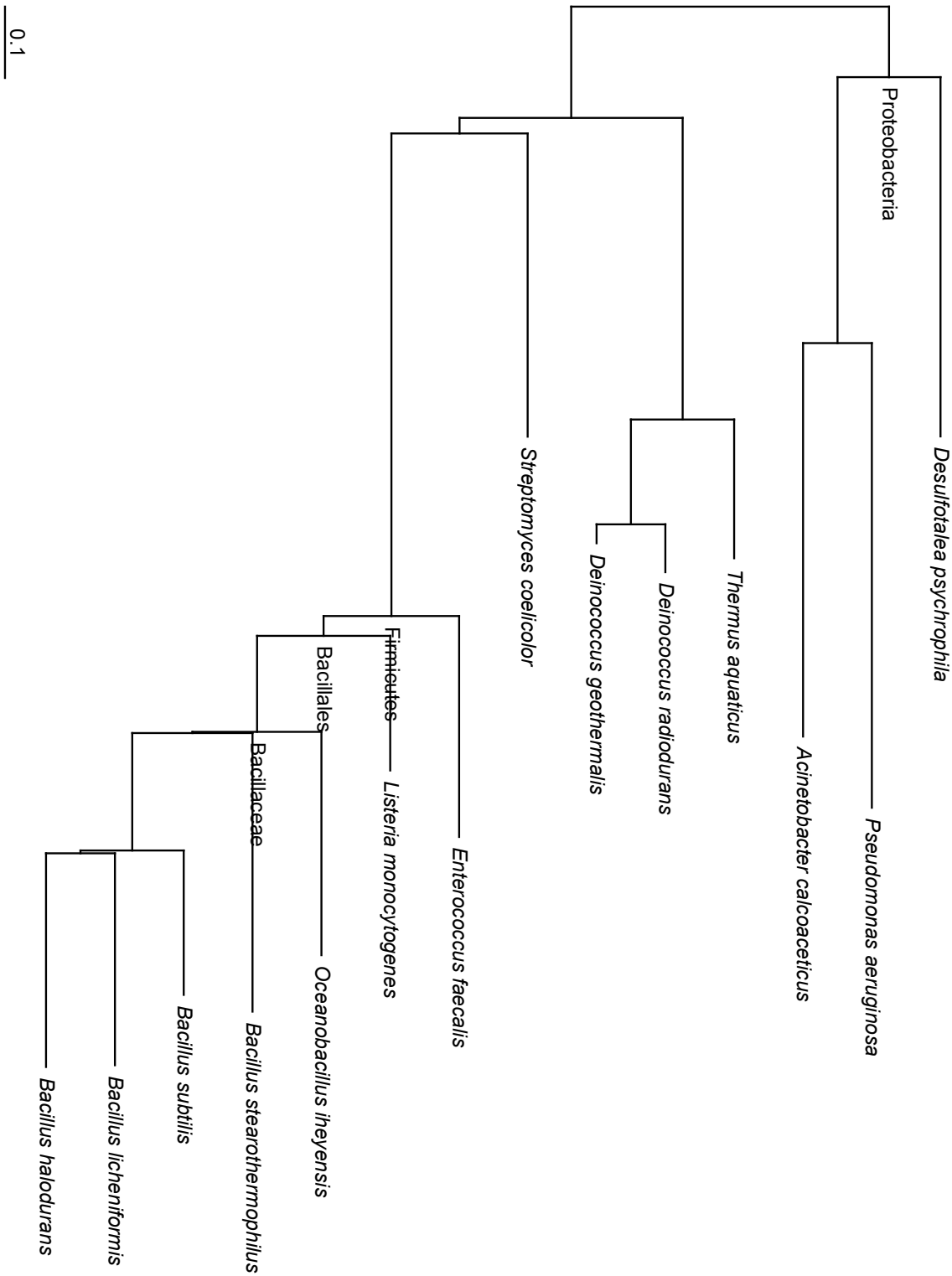


Figure 4.T.r1.s2.c.p.bacteria: Round 1 subset 2 of final tree, Bacteria only shown, phylogram

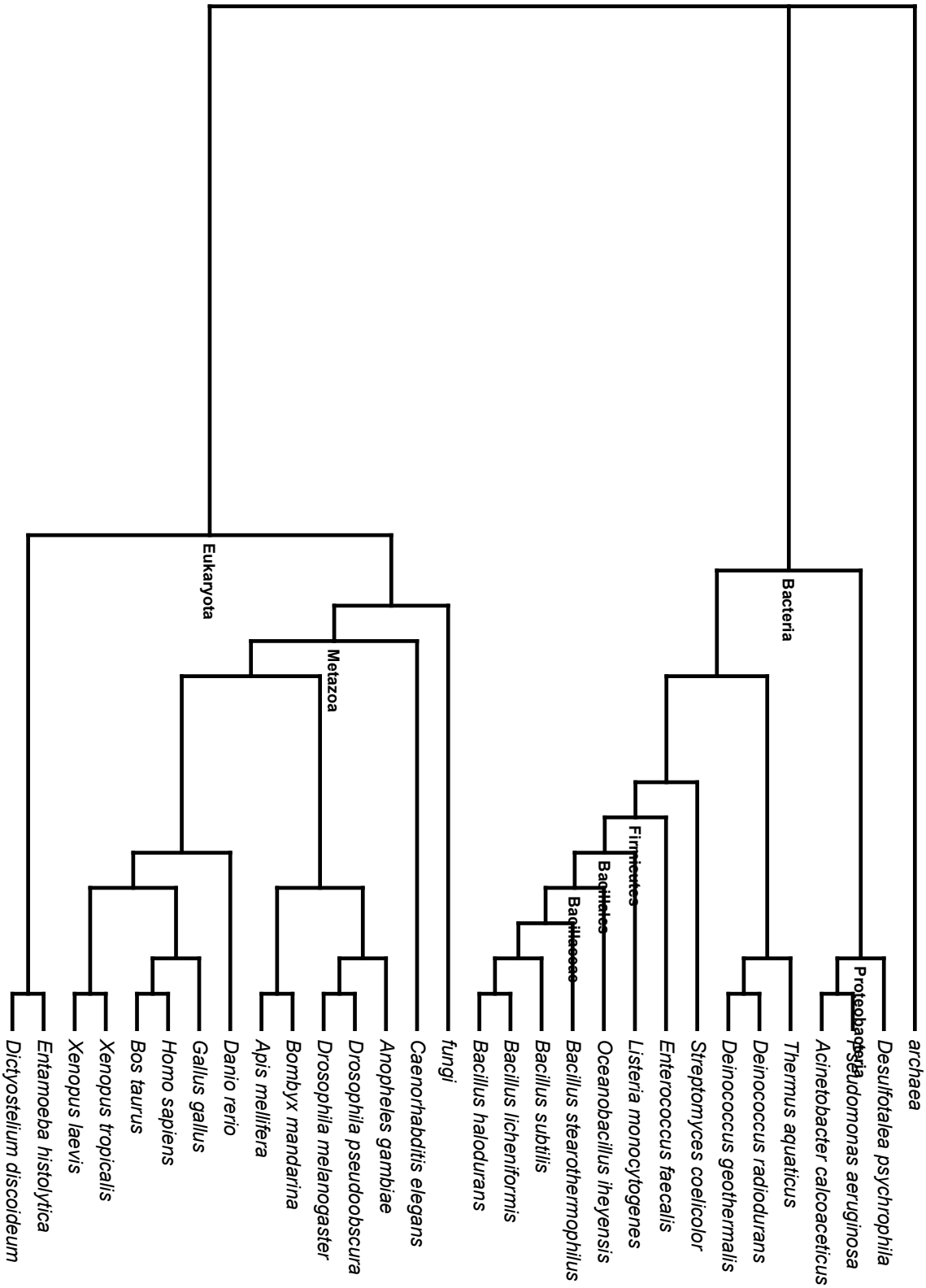


Figure 4.T.r1.s2.c.c: Round 1 subset 2 of final tree, cladogram

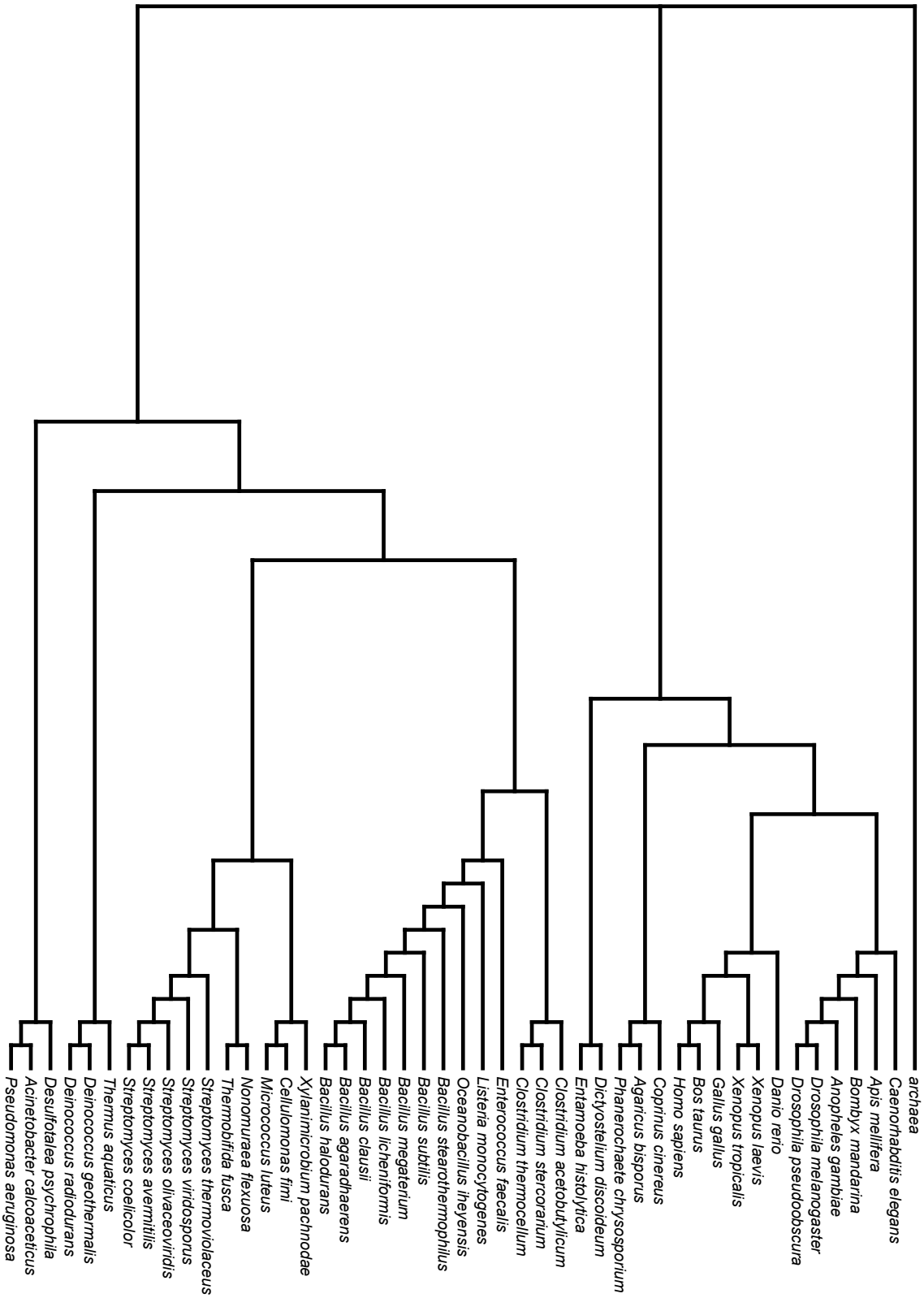


Figure 4.T.r1.s2.1: Round 1 subset 2, original (tree 1) arrangement, cladogram

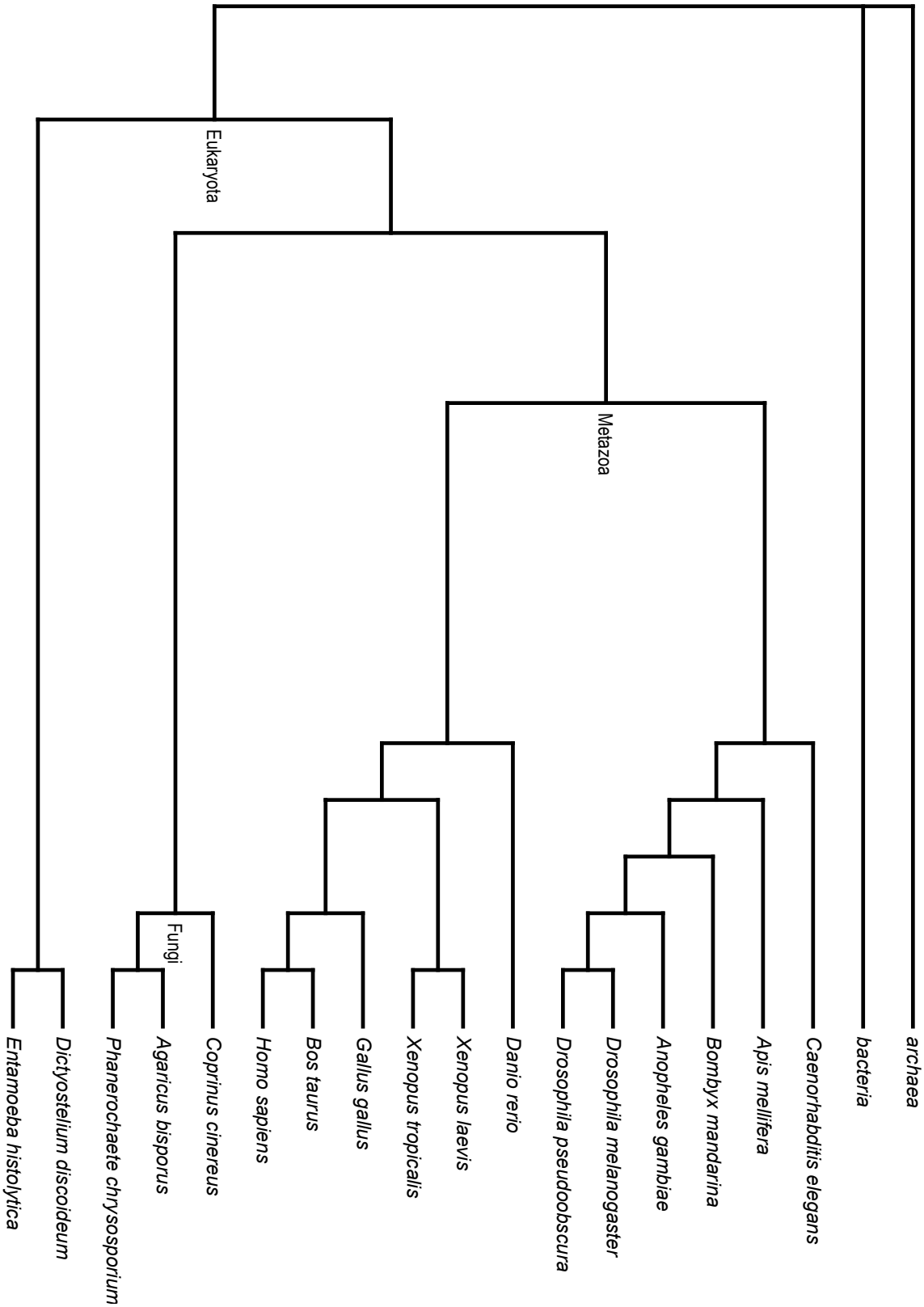


Figure 4.T.r1.s2.1.eukaryota: Round 1 subset 2, original (tree 1) arrangement, Eukaryota only shown, cladogram

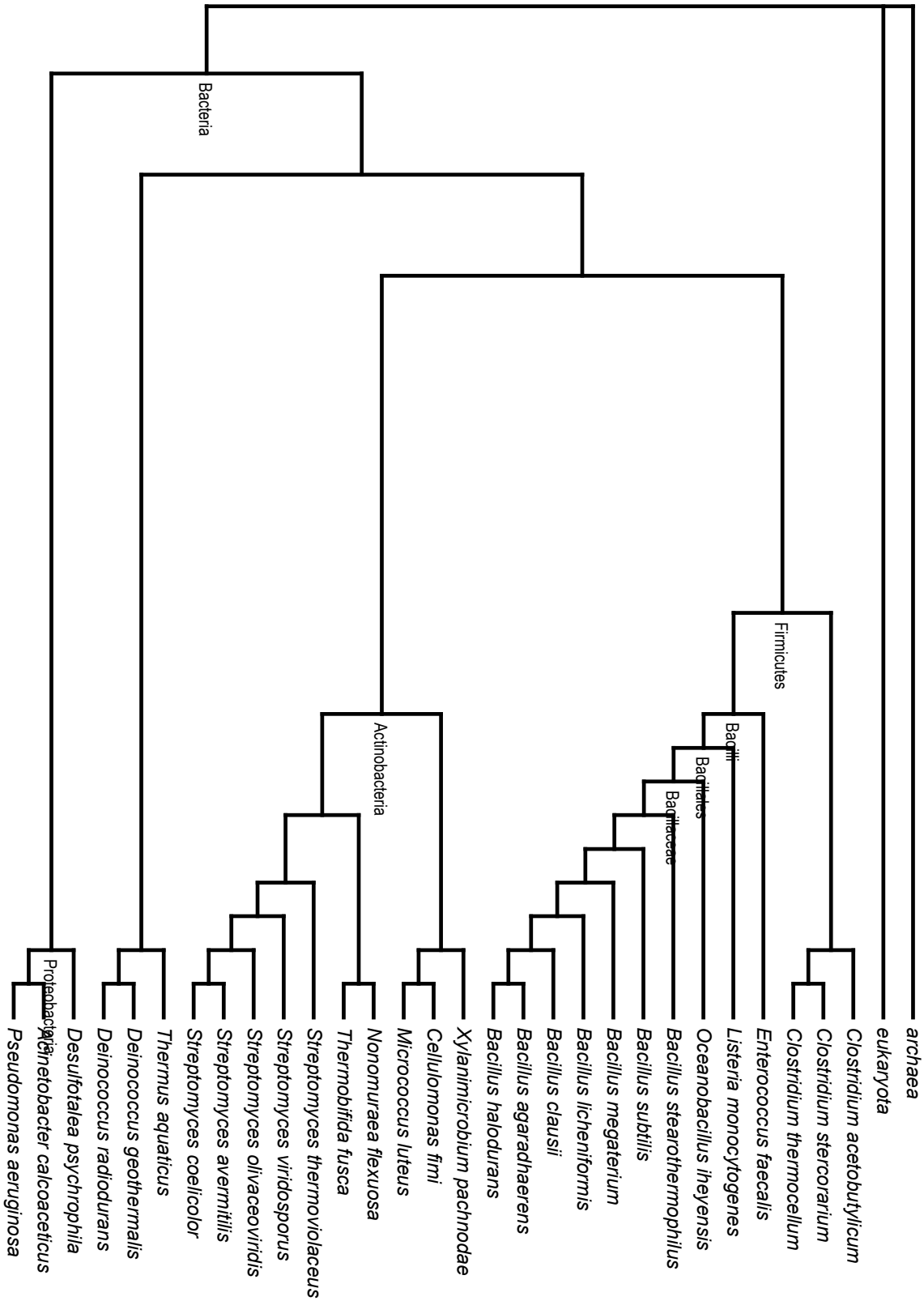


Figure 4.T.r1.s2.1.bacteria: Round 1 subset 2, original (tree 1) arrangement, Bacteria only shown, cladogram

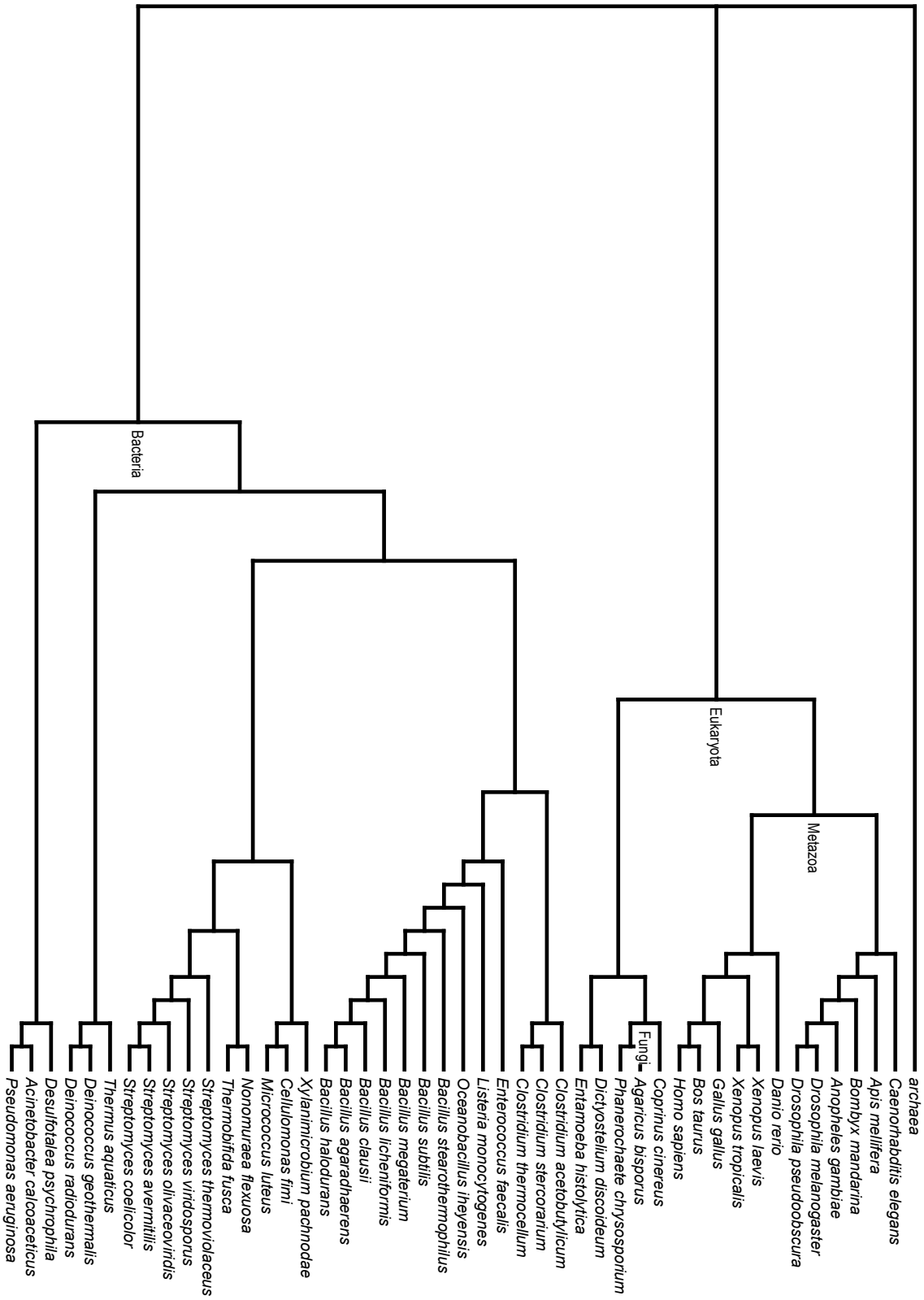


Figure 4.T.r1.s2.12: Round 1 subset 2, Tree 12 arrangement, cladogram

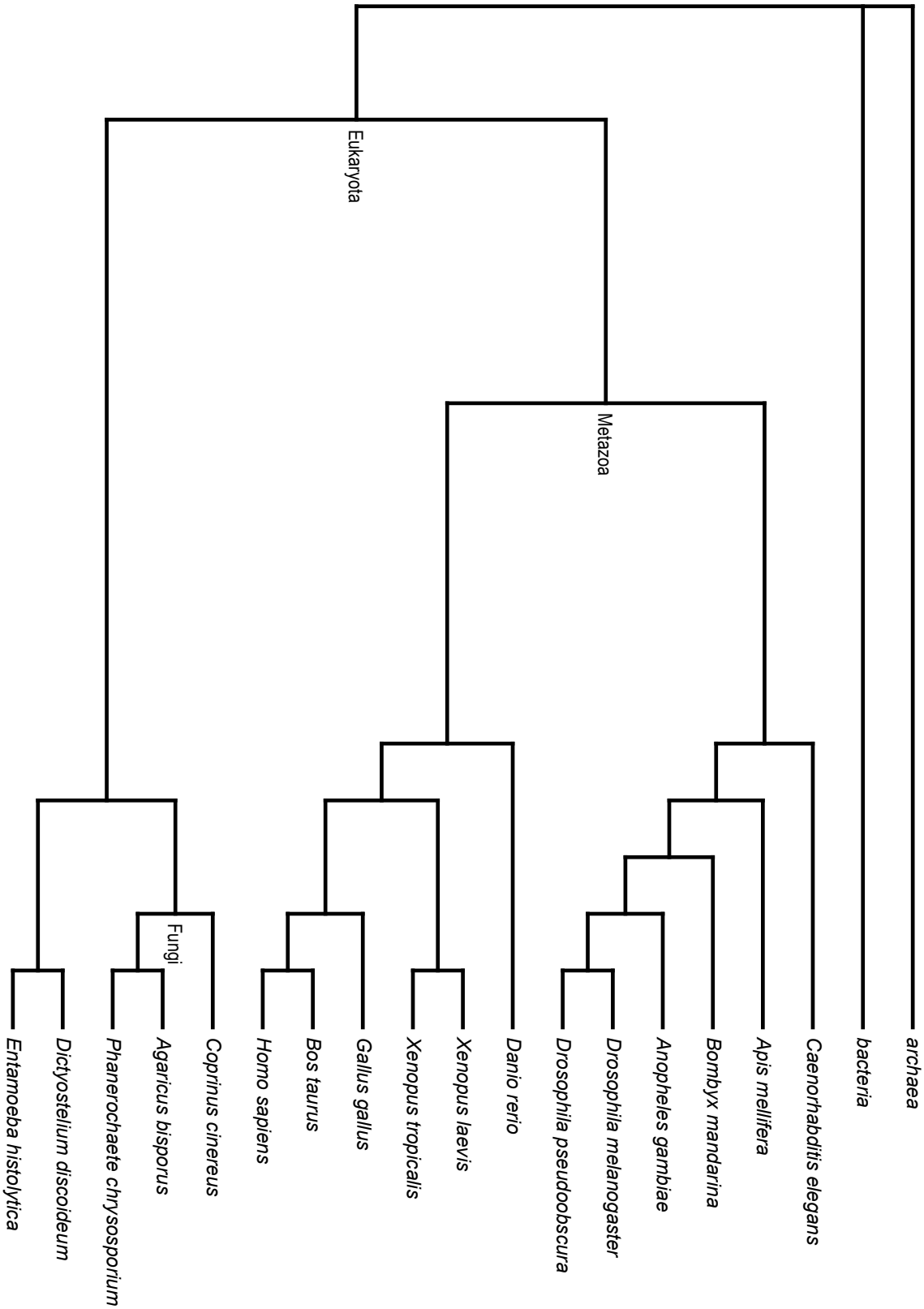


Figure 4.T.r1.s2.12.eukaryota: Round 1 subset 2, Tree 12 arrangement, Eukaryota only shown, cladogram

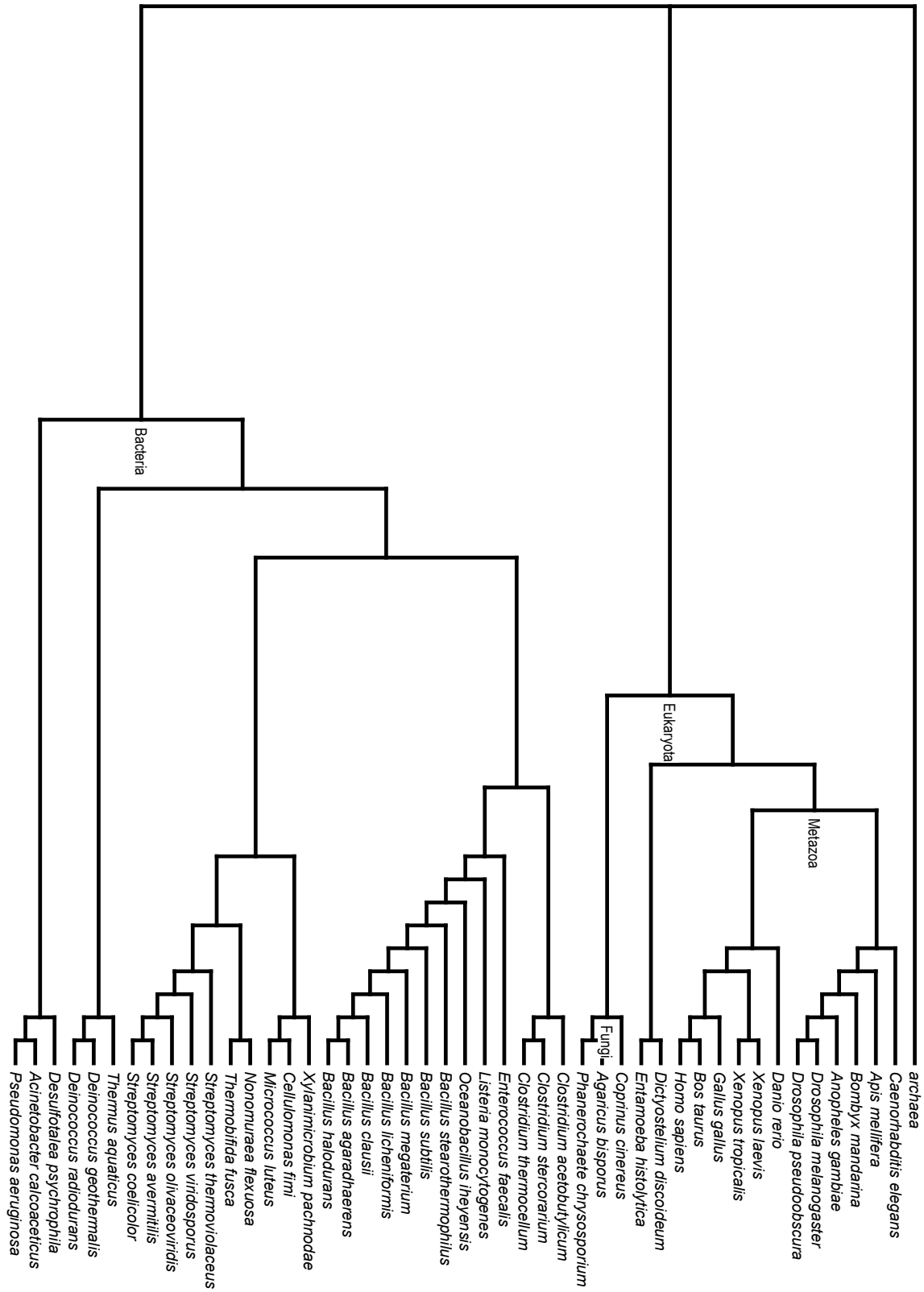


Figure 4.T.r1.s2.13: Round 1 subset 2, Tree 13 arrangement, cladogram

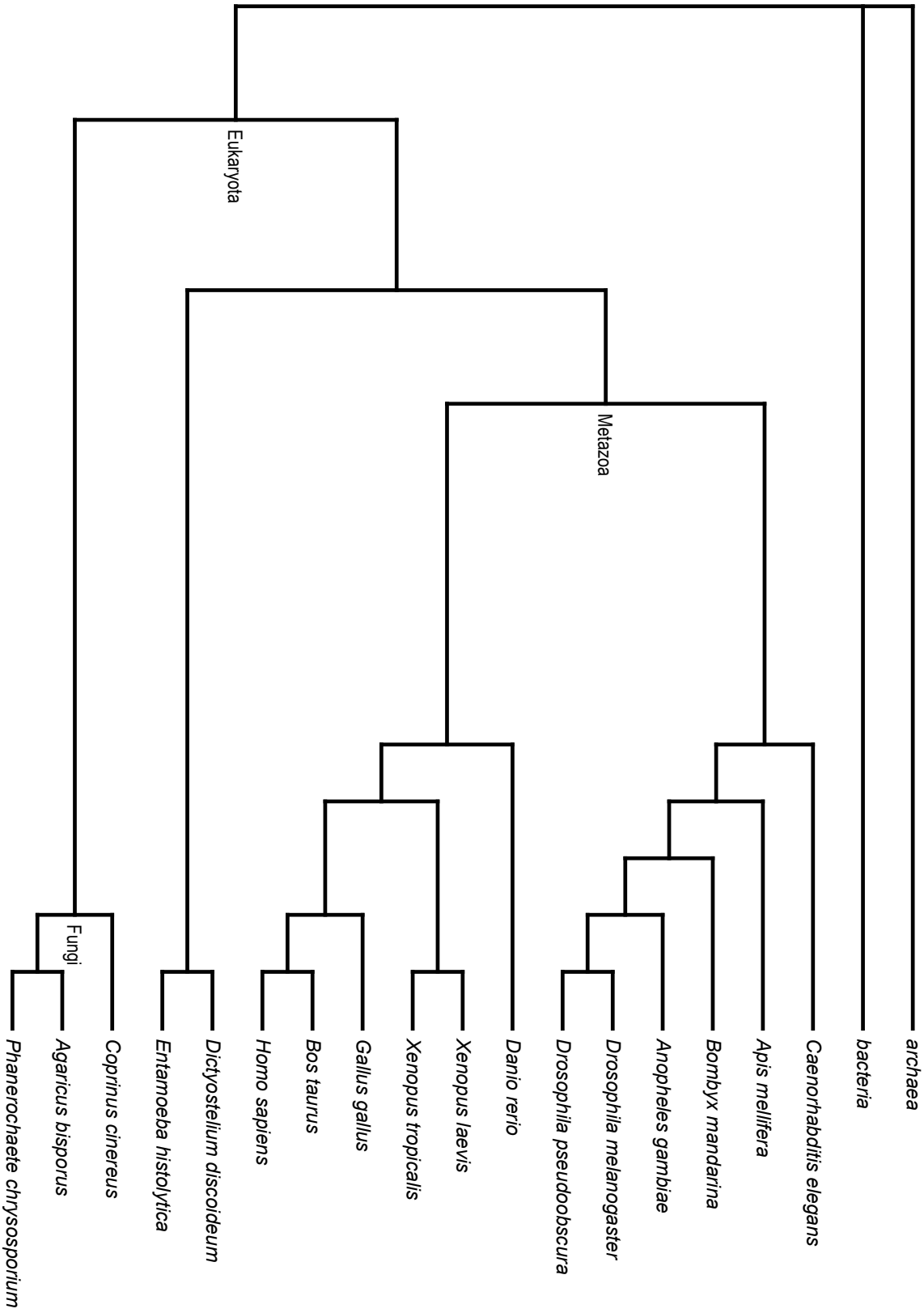


Figure 4.T.r1.s2.13.eukaryota: Round 1 subset 2, Tree 13 arrangement, Eukaryota only shown, cladogram

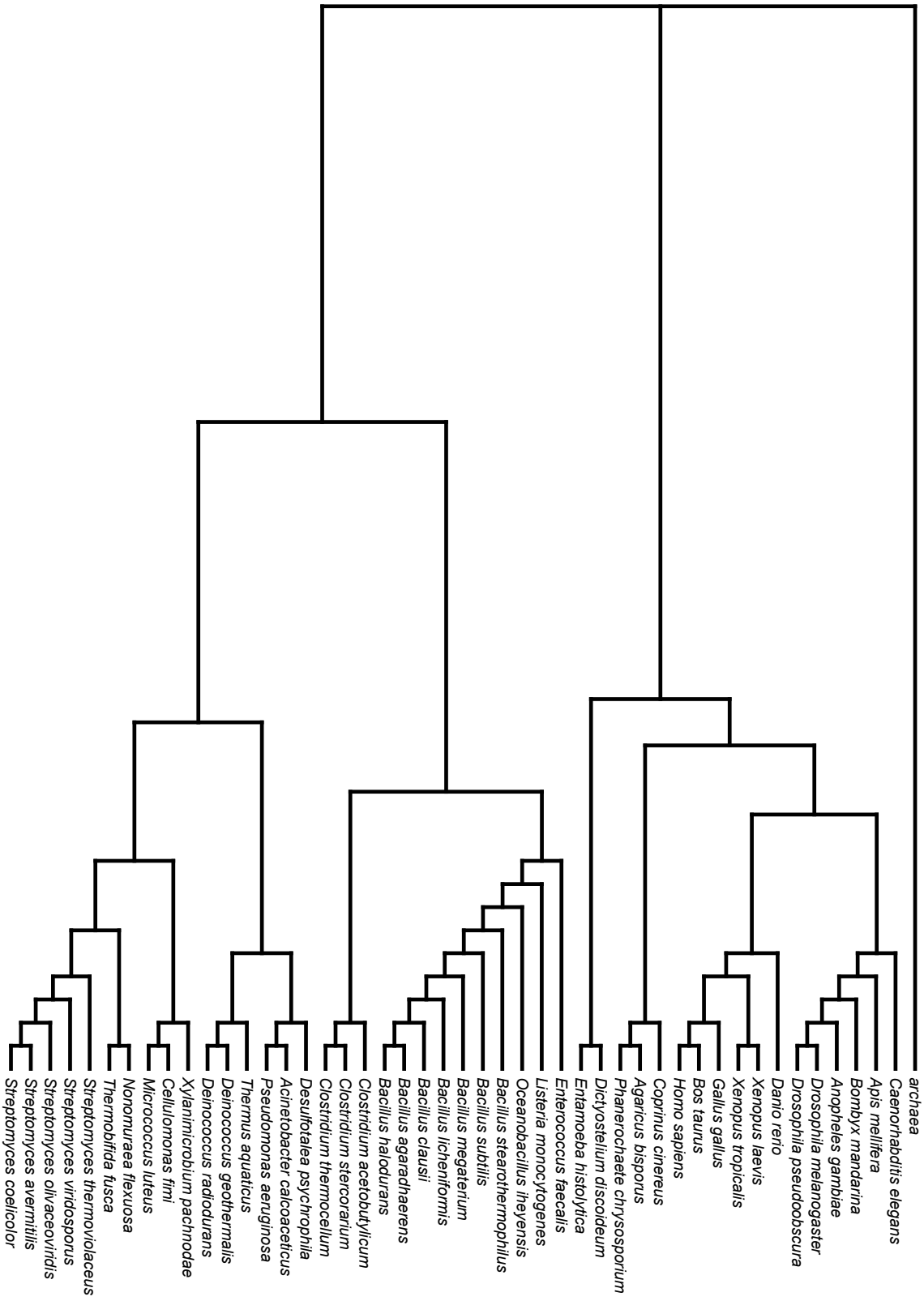


Figure 4.T.r1.s2.15: Round 1 subset 2, Tree 15 arrangement, cladogram

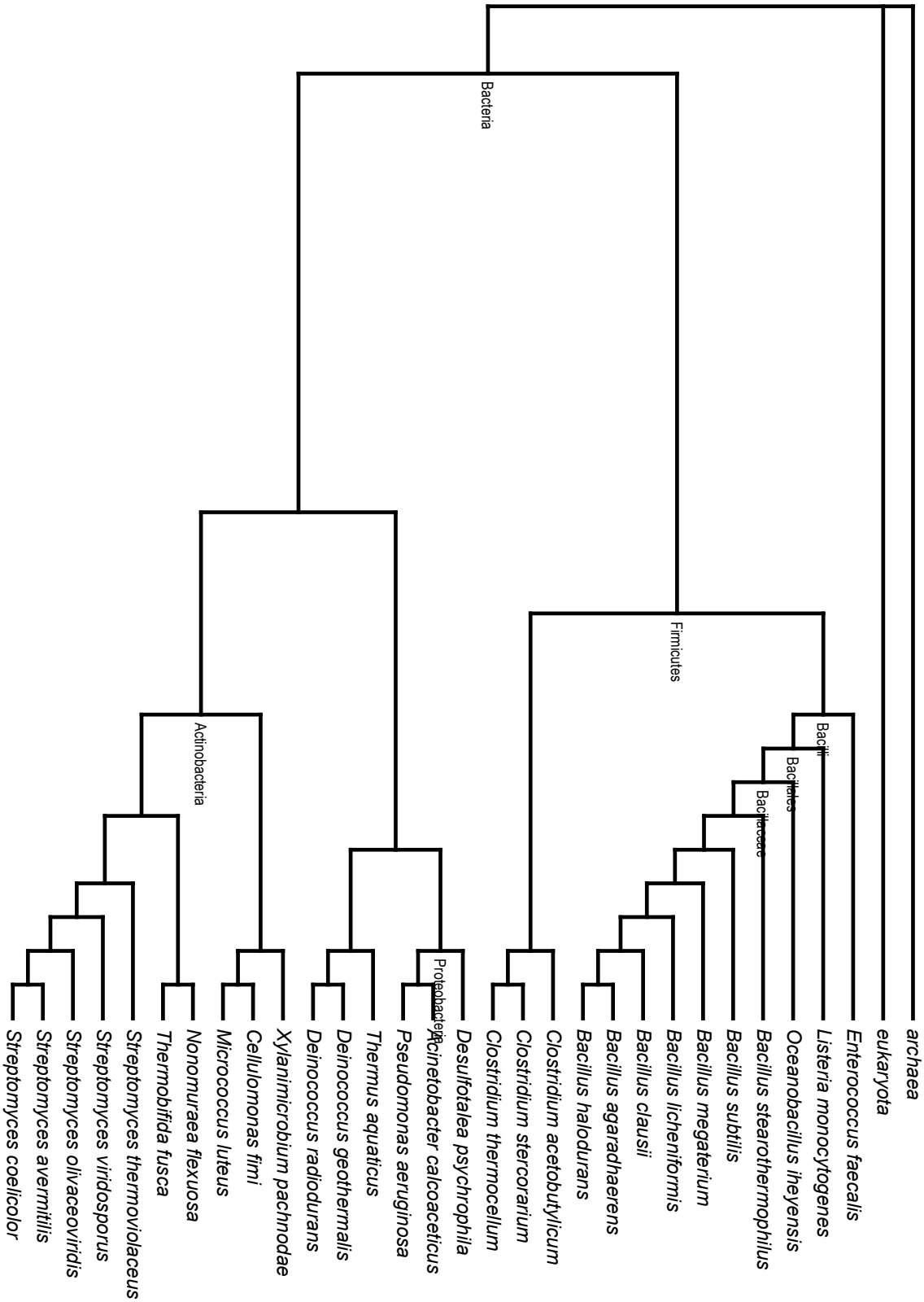


Figure 4.T.r1.s2.15.bacteria: Round 1 subset 2, Tree 15 arrangement, Bacteria only shown, cladogram

The overall conclusions from the above are:

- From 12 and 13, that *D. discoideum* and *E. histolytica* are not closer to fungi or metazoa
- From 15, that the existing arrangement of bacterial species was preferable.

Subset 5: Some Eukaryota

For subset 5, runs were done with 5868 amino acids from 18 proteins (counting ADH1 as 1 protein), with 200000 generations (2000 samples) and a "burnin" for sump of 1000. The log probabilities were as follows:

Phylogeny Tested	Arithmetic Mean	Harmonic Mean
Original (1)	-182,428.91	-182,516.27
2	-182,479.96	-182,539.93
3	-182,496.69	-182,570.28
4	**-175,595.86**	**-175,865.64**

(Note that, as stated on page 230, the final tree is as per tree arrangement (hypothesis) 4.) The Metazoa species are in the following groupings:

- By the "classical" definition (tree arrangements 2, 3, and 4):
  - Coelomata: See Figure 4.T.r1.s5.c.p.eukaryota, on page 224, for information on what species are in Deuterostomia and Protostomia.
  - Acoelomata: *Schistosoma japonicum*, *Schistosoma mansoni*
  - Pseudocoelomata: See Figure 4.T.r1.s5.c.p.eukaryota, on page 224, for information on what species are in Pseudocoelomata (abbreviated "Pseudocoel.").
- By the Ecdysozoa and Lophotrochozoa definition (tree arrangement 1):
  - Deuterostomia: As per the "classical" definition.

- Ecdysozoa plus Lophotrochozoa branch:
  - Ecdysozoa: Classical “Protostomia” (see above) plus Pseudocoelomata (see above)
  - Lophotrochozoa: Classical “Acoelomata”, *Schistosoma japonicum* and *Schistosoma mansoni*

The trees are on pages 223-229.

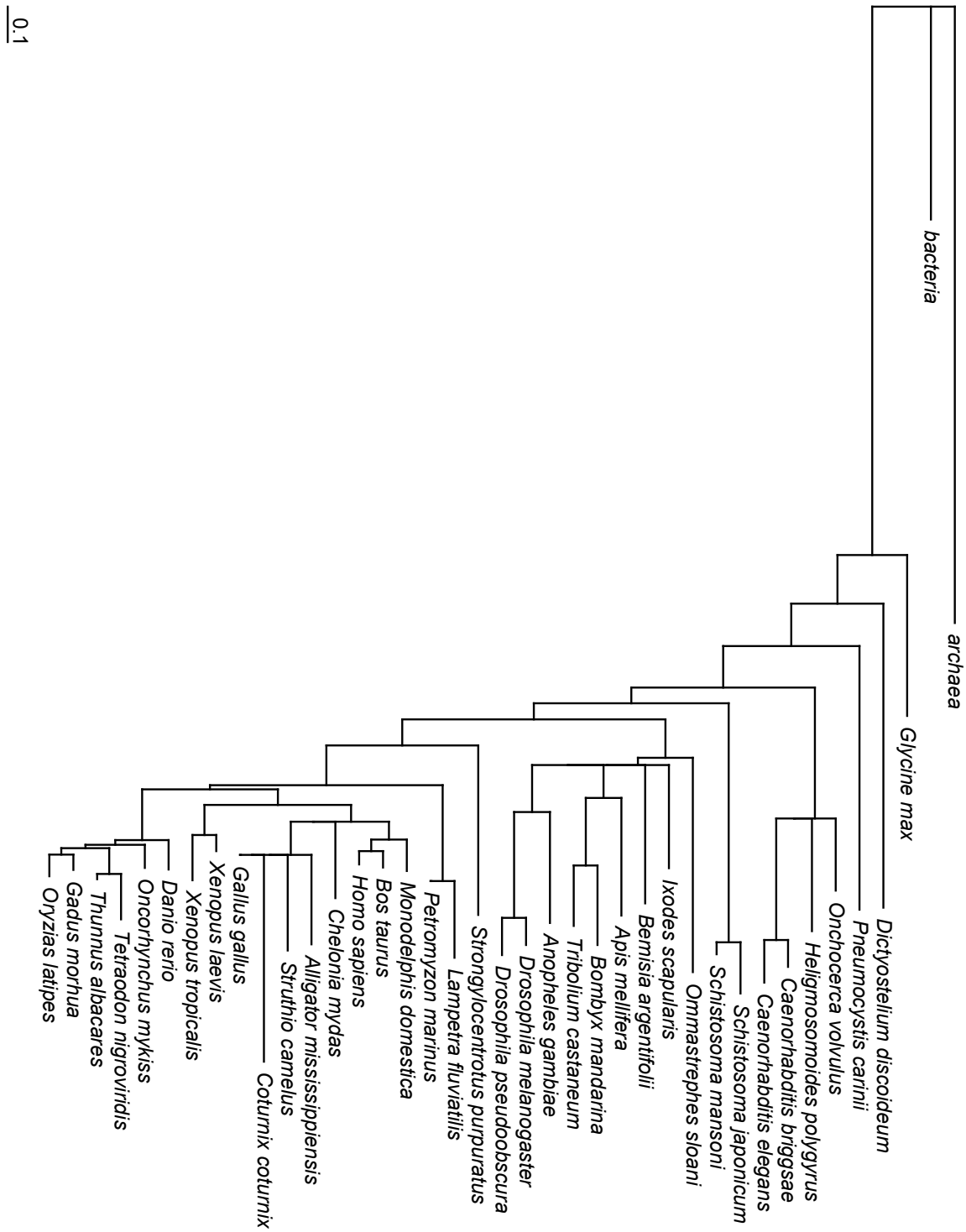


Figure 4.T.r1.s5.c.p: Round 1 subset 5 of final tree, phylogram

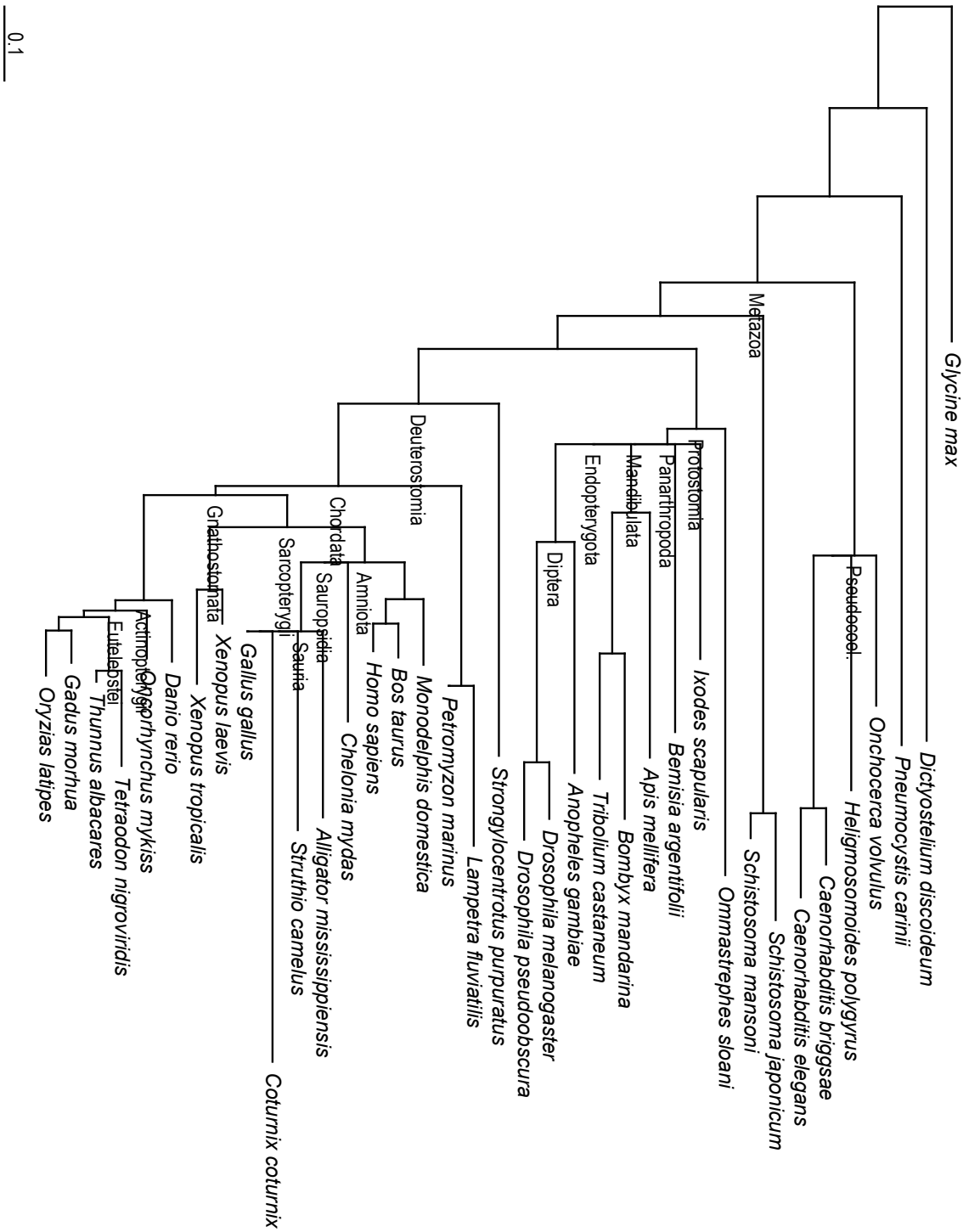


Figure 4.T.r1.s5.c.p.eukaryota: Round 1 subset 5 of final tree, Eukaryota only shown, phylogram

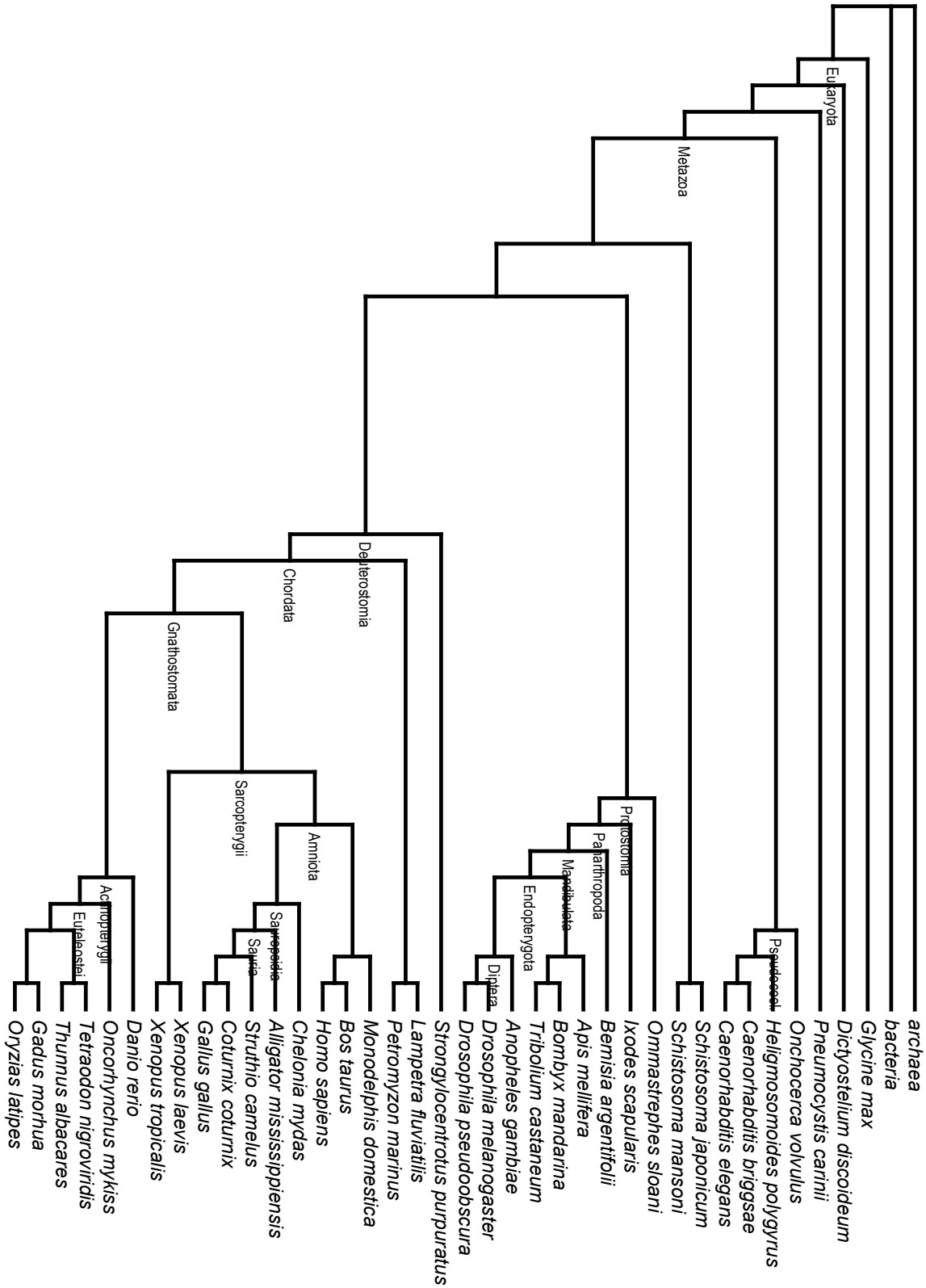


Figure 4.T.r1.s5.c.c: Round 1 subset 5 of final tree, cladogram

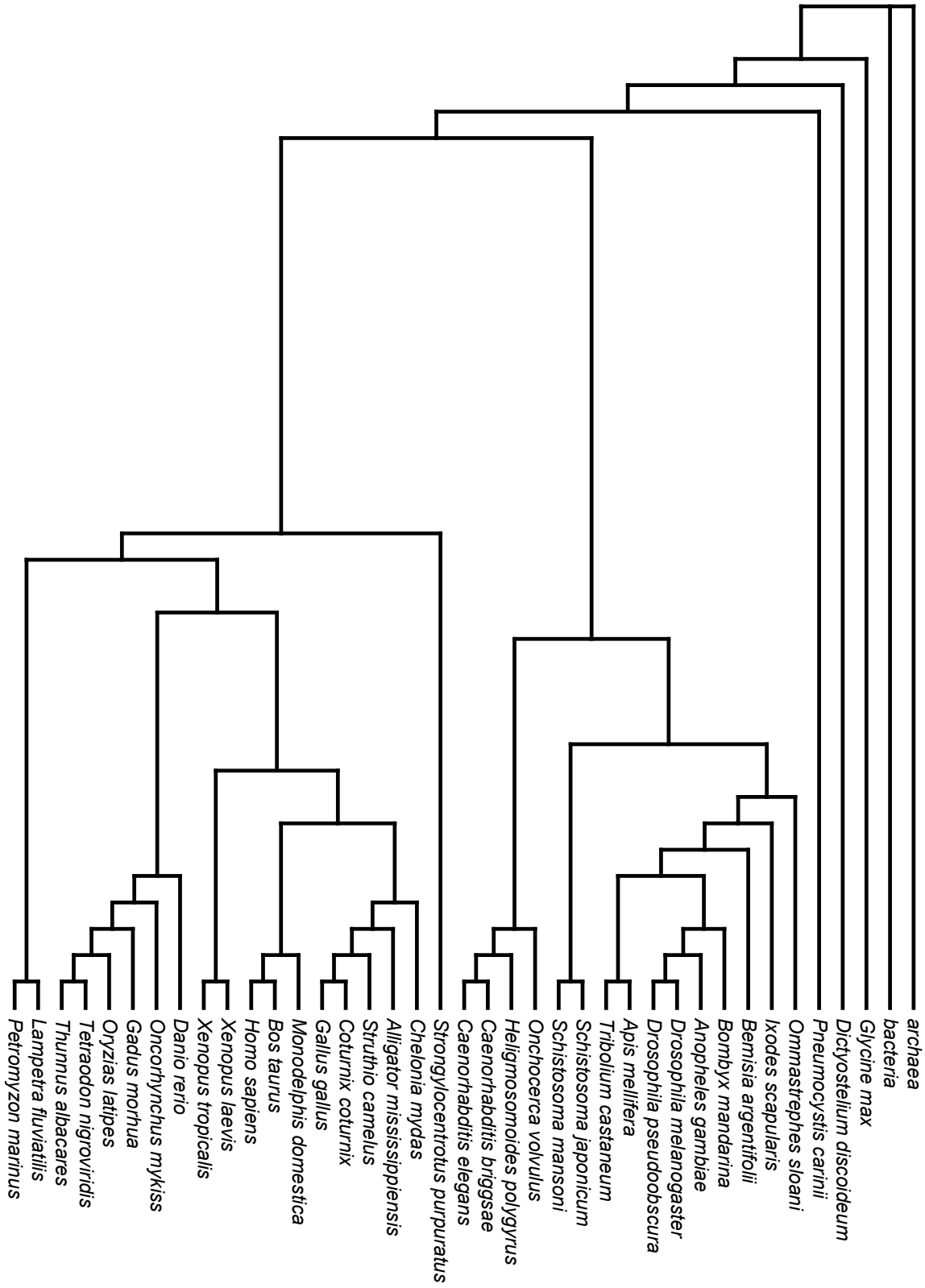


Figure 4.T.r1.s5.1: Round 1 subset 5, original (tree 1) arrangement, cladogram

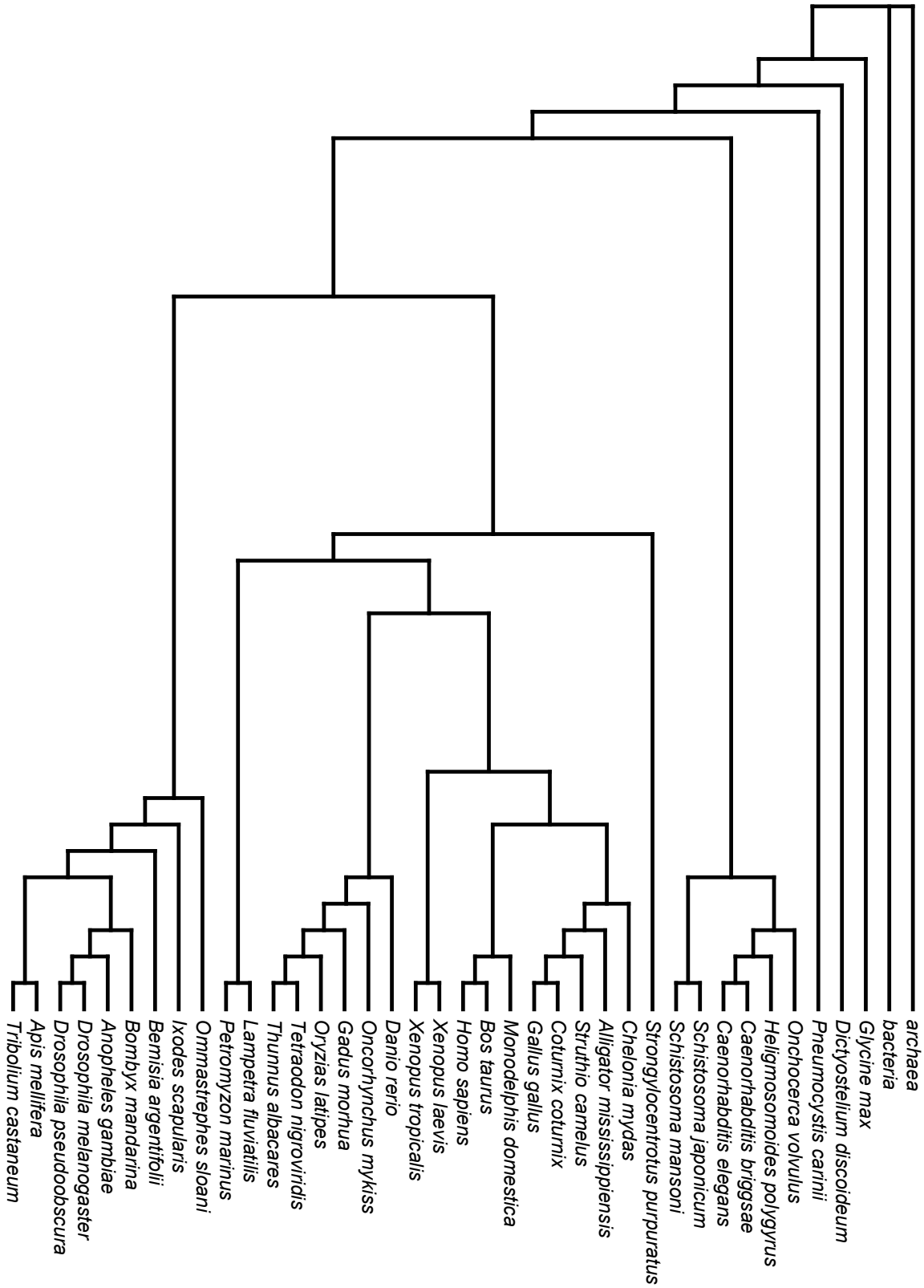


Figure 4.T.r1.s5.2: Round 1 subset 5, tree 2 arrangement, cladogram

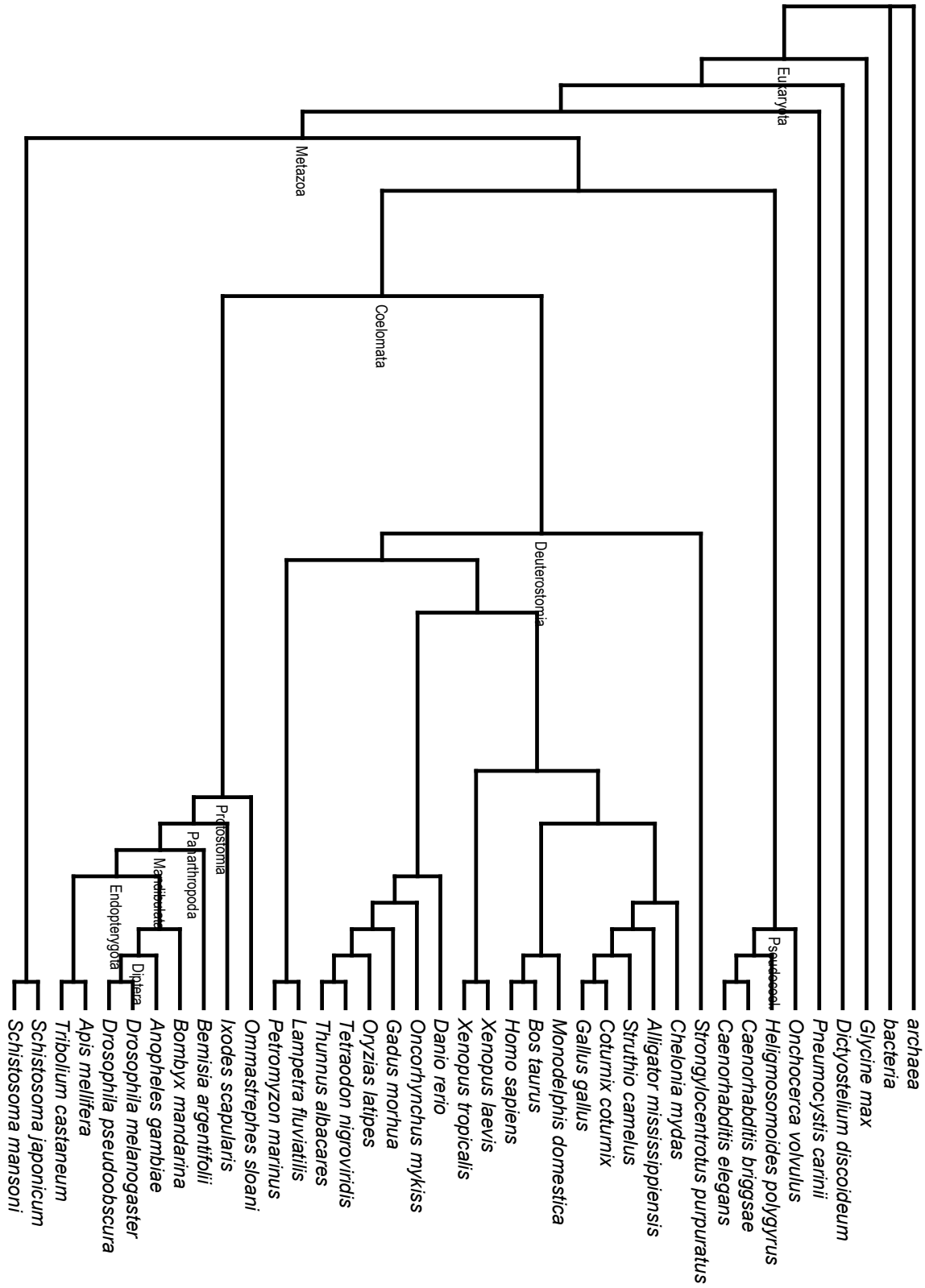


Figure 4.T.r1.s5.3: Round 1 subset 5, tree 3 arrangement, cladogram

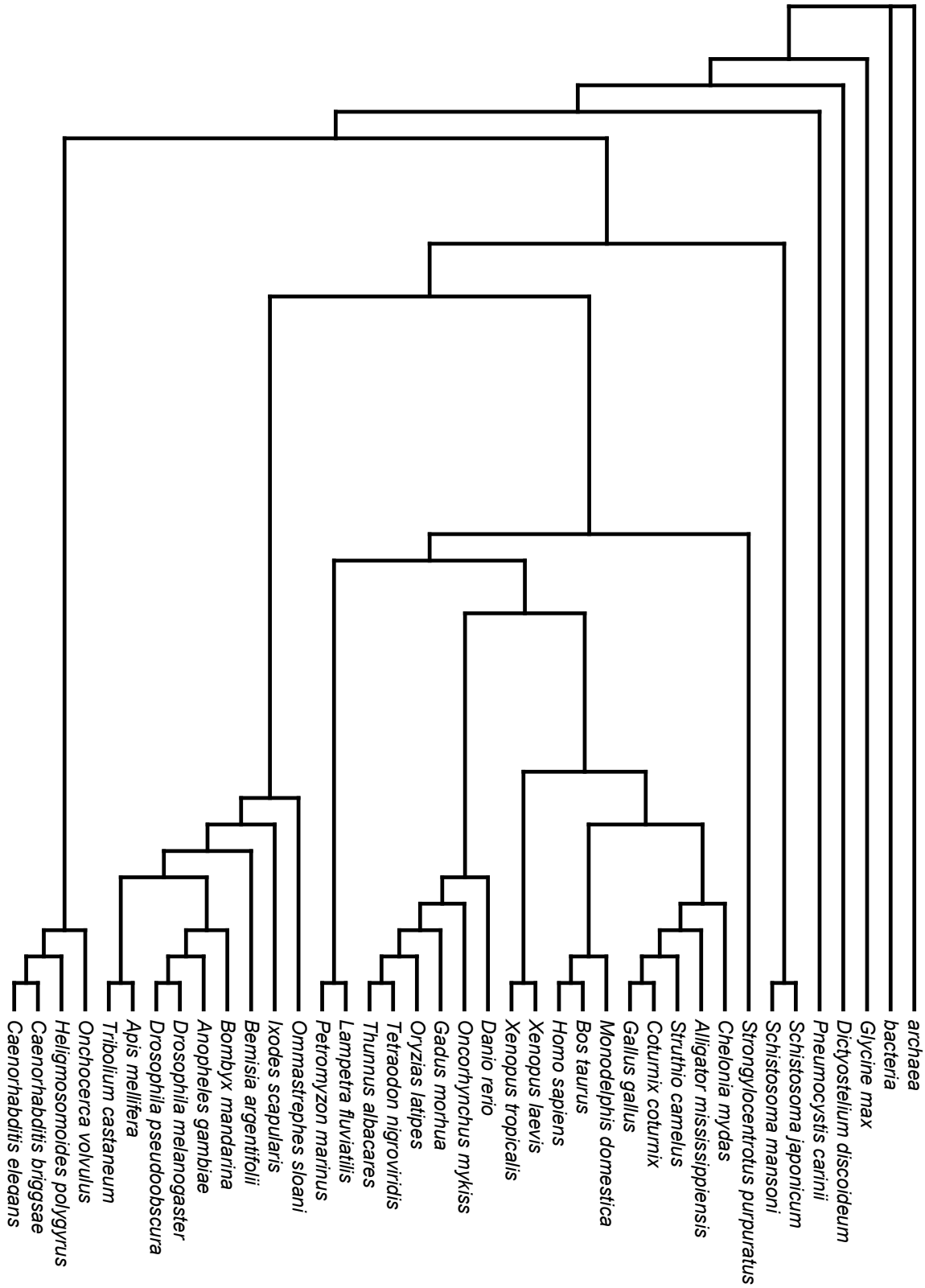


Figure 4.T.r1.s5.4: Round 1 subset 4, tree 4 arrangement, cladogram

Together with the results from subset 6 (see "Subset 6: Some Eukaryota (and others)", on page 231), the results from subset 5 indicate that a more "classical"[449] division than the original of Metazoa into Coelomata (e.g., Vertebrata), Pseudocoelomata (e.g., Nematoda), and Acoelomata (e.g., *Schistosoma*), is preferable[450]. Moreover, from tree 4 having the best probabilities from subset 5, Pseudocoelomata is indicated as branching prior to Coelomata and Acoelomata (a finding interestingly at odds with the usual assumption of Acoelomata as the "simplest" organisms and therefore branching first (Philippe, Lartillot, & Brinkmann 2005)).

---

[449] This is as per the NCBI taxonomy.

[450]This finding is unexpected, given the strong earlier evidence otherwise (Philippe, Lartillot, & Brinkmann 2005; Ruiz-Trillo *et al.* 1999; Telford, Wise, & Gowri-Shankar 2005). Long-branch attraction artifacts, as noted in earlier work (Philippe, Lartillot, & Brinkmann 2005), are a potential problem:

- given that we were not able to use the covarion option in MrBayes (see page 99, footnote 200), especially given the prior research on the subject taking into account rRNA's stem-loop structure (Telford, Wise, & Gowri-Shankar 2005);
- the long branch length (see Figure 4.T.r1.s5.c.p.eukaryota, on page 224) for *Schistosoma* (Acoelomata), although it is notable that Acoelomata is not the outermost group;
- that only two species in Acoelomata were available, both in the same genus, *Schistosoma*;
- the long branch length for *C. elegans*, although that for *Onchocerca volvulus* is not as long so should help correct for any problems (Anderson & Swofford 2004; Gibb *et al.* 2007; Graham, Olmstead, & Barrett 2002; Moreira, Lopez-Garcia, & Vickerman 2004);
- that the species in Pseudocoelomata are all Nematoda (and, indeed, are all Chromadorea).

Given the strength of the probability differences, however, we concluded that we were unlikely to find alternative trees with reasonable justifications, so decided to use the arrangement as per tree 4, despite prior evidence (and our earlier hypothesis, given said evidence) otherwise. This is an area for further exploration.

Subset 6: Some Eukaryota (and others)

For subset 6, runs were done with 9878 amino acids from 25 proteins (counting ADH1 as 1 protein), with 200000 generations (2000 samples). The log probability results from subset 6, for the indicated burnin periods, were as follows:

Phylogeny Tested	Burnin = 1000		Burnin = 1500	
	Arith. M.	Harmon. M.	Arith. M.	Harmon. M.
1 (original):	-207,369.05	-229,262.13	-207,369.05	-207,754.65
2|3|4	**-190,294.90**	**-191,183.08**	**-190,294.90**	**-190,897.05**
12:	-217,851.52	**-226,330.76**	-217,851.52	-218,739.70
13:	**-172,611.54**	-230,644.28	**-172,611.54**	**-191,096.72**

The Metazoa species are in the following groupings:

- By the "classical" definition (tree arrangements 2, 3, and 4):
  - Coelomata: See Figure 4.T.r1.s6.c.p.eukaryota, on page 233, for information on what species are in Deuterostomia and Protostomia.
  - Pseudocoelomata: *C. briggsae* and *C. elegans*
- By the Ecdysozoa and Lophotrochozoa definition (tree arrangement 1):
  - Deuterostomia: As per the "classical" definition.
  - Ecdysozoa: Classical "Protostomia" (see above) plus Pseudocoelomata (see above)

The species rearranged for 1 versus 12 versus 13 is *D. discoideum*. The final (as per 2|3|4) and original (tree 1) configurations of subset 6's species are shown on the following pages.

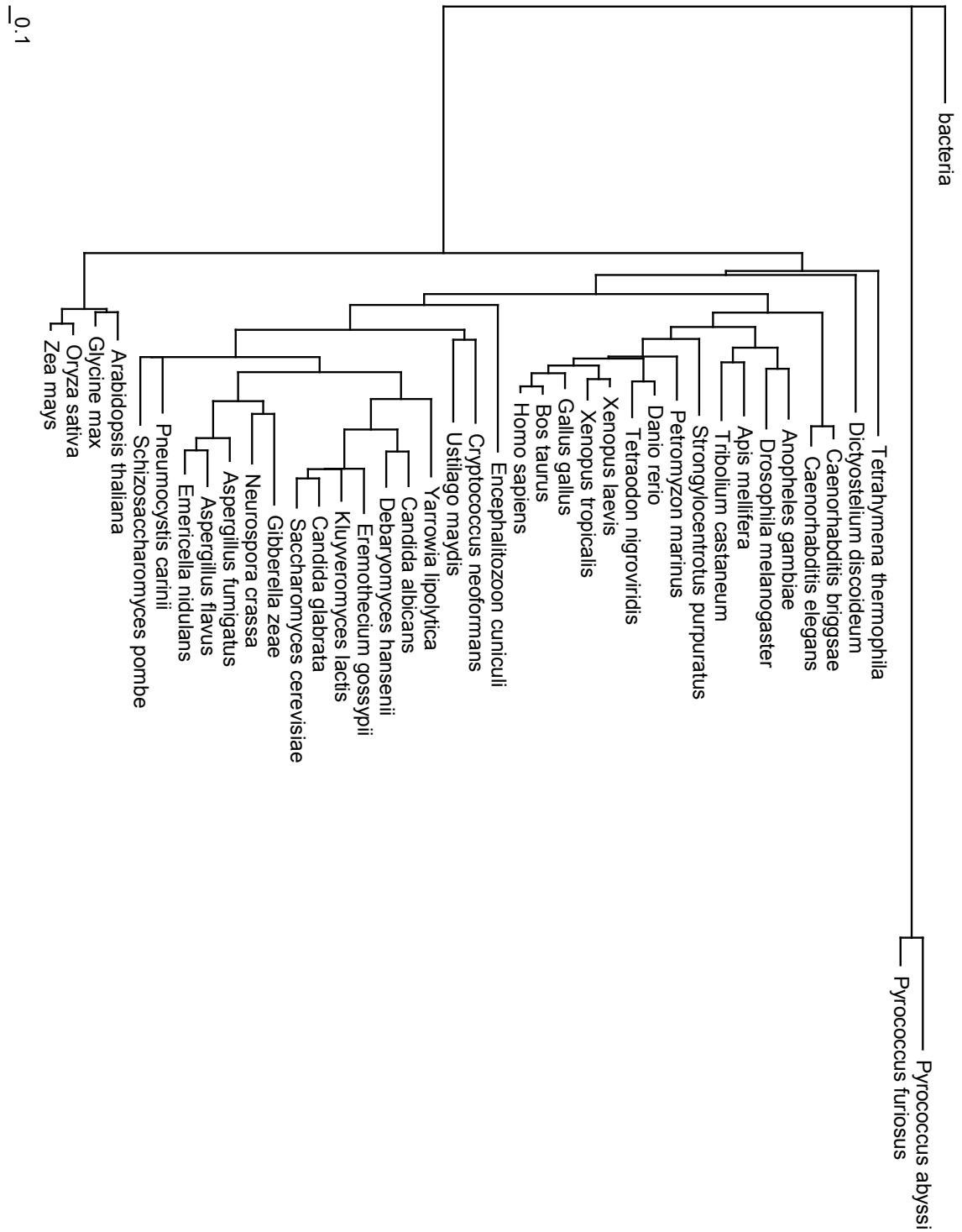


Figure 4.T.r1.s6.c.p: Round 1 subset 6 of final tree, phylogram

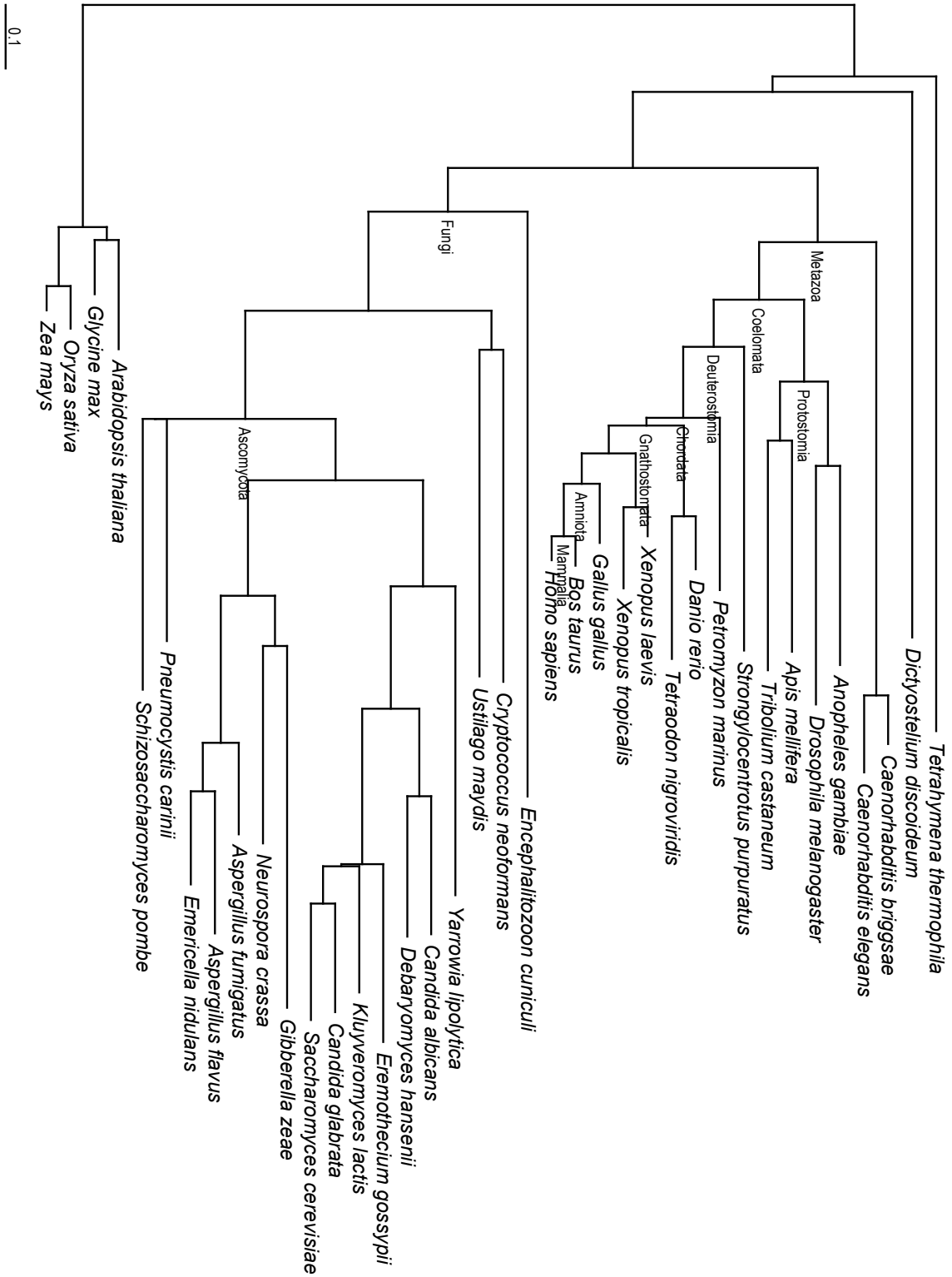


Figure 4.T.r1.s6.c.p.eukaryota: Round 1 subset 6 of final tree, Eukaryota only shown, phylogram

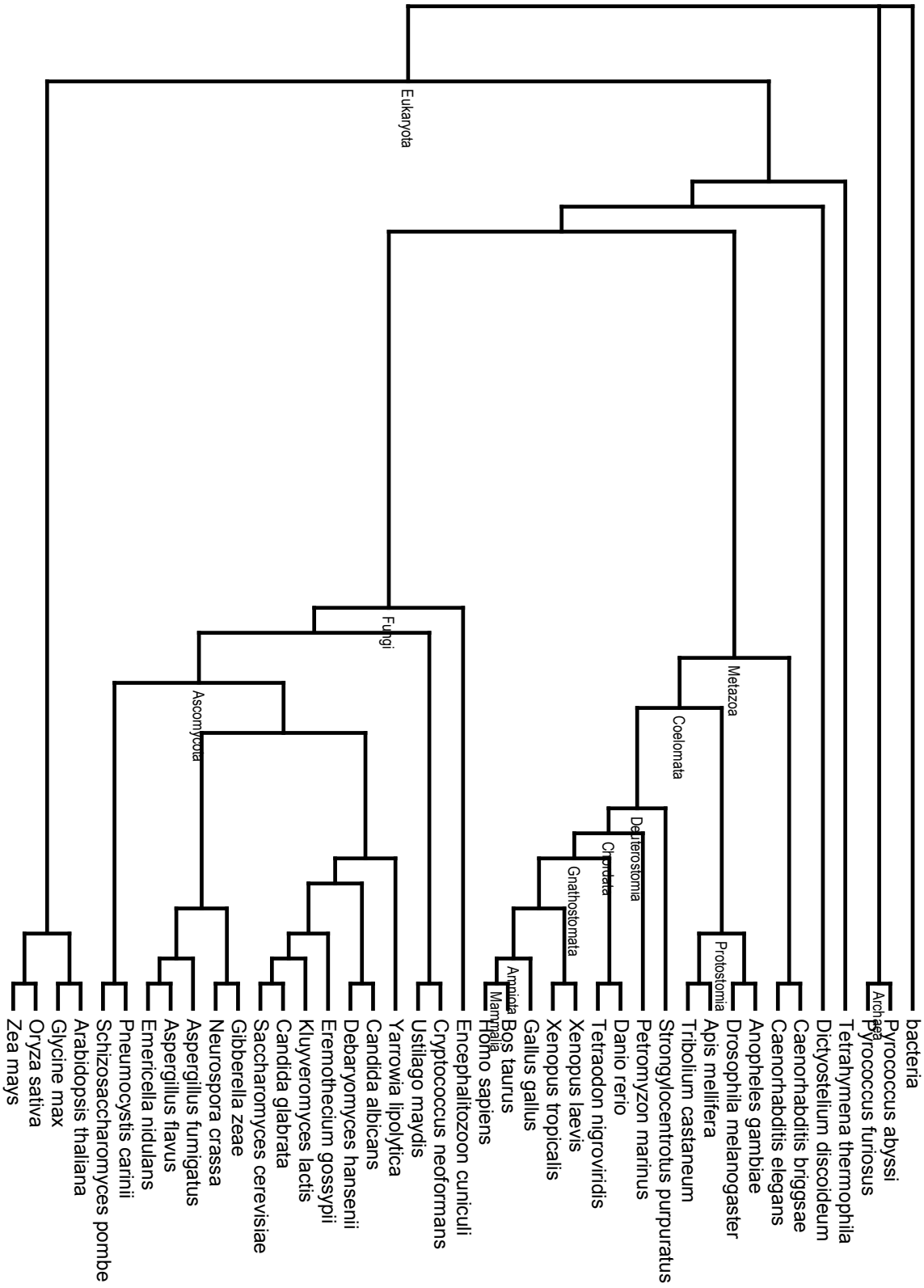


Figure 4.T.r1.s6.c.c: Round 1 subset 6 of final tree, cladogram

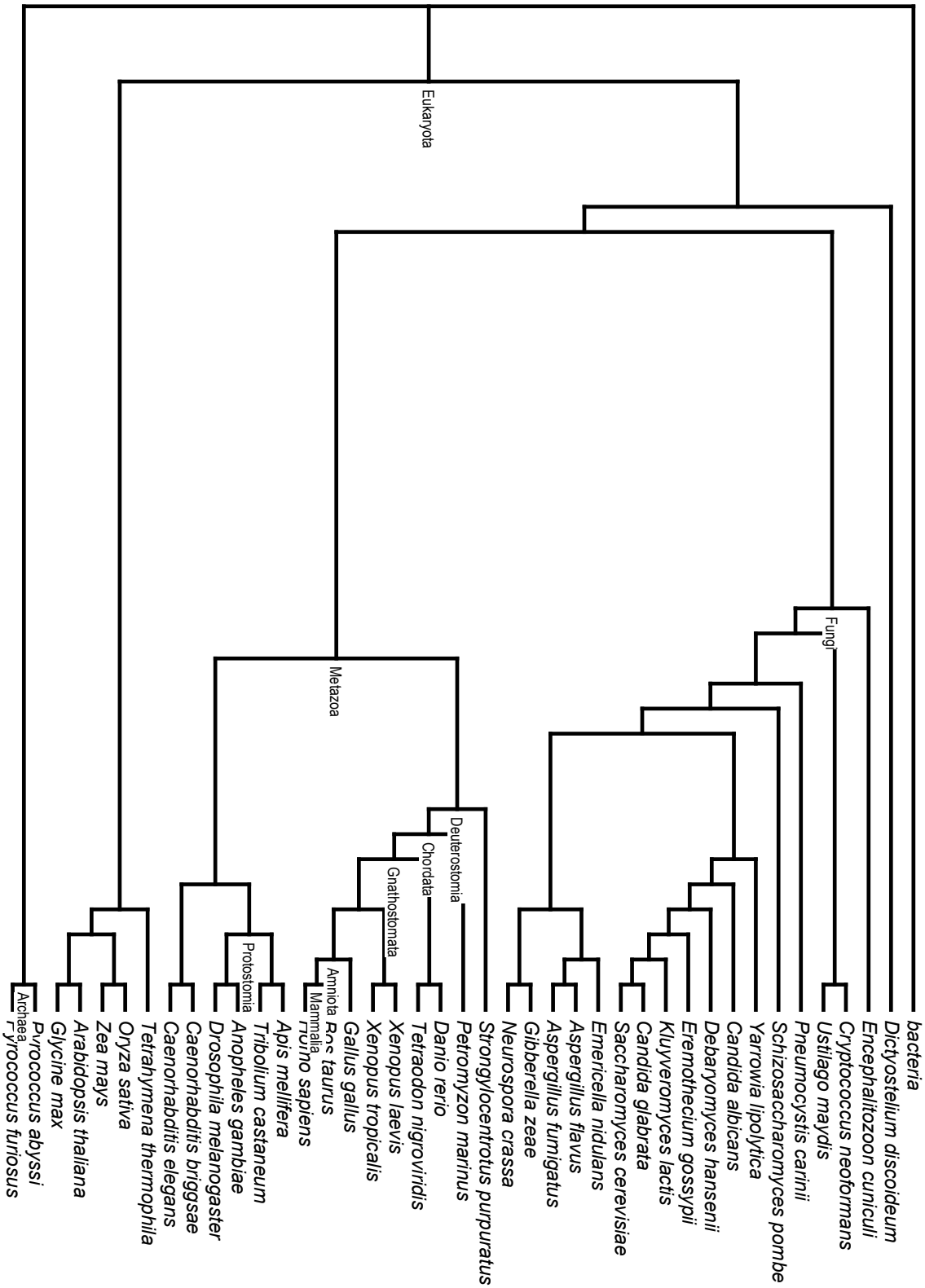


Figure 4.T.r1.s6.1: Round 1 subset 6, original (tree 1) arrangement, cladogram

For other trees with subset 6, the rearrangements in the same phylogeny variant numbers are the same as with subsets 2 and 5 (and all others in round 1); therefore, please see “Appendix L: Tree files available”, on page 394, for:

- round1.subset.6.usertree.2.phy - identical to:
  - round1.subset.6.usertree.3.phy
  - round1.subset.6.usertree.4.phy
- round1.subset.6.usertree.12.phy
- round1.subset.6.usertree.13.phy

Subset 1: Some Proteobacteria, Eukaryota

For subset 1, runs were done with 2556 amino acids[451], from 7 proteins. The runs were for 200000 generations (2000 samples), with a burnin of 1000. The final status of the species[452] in this subset can be seen in the trees from pages 237 to 240. The log probabilities[453] were as follows:

Phylogeny Tested	Arith. M.	Harmon. M.
1 (original):	**-38,059.52**	-102,041.25
12:	-44,687.40	**-59,405.16**
13:	-55,125.89	-60,202.35

The species rearranged for 1 versus 12 versus 13 is *D. discoideum*.

[451] As well as the problems noted below with the outgrouping, it is likely that this is too few amino acids for the number of species.

[452] The following species in the original of subset 1 are not in the current tree: *Burkholderia fungorum, Haemophilus ducreyi, Histophilus somni, Mannheimia succiniciproducens, Pasteurella multocida, Pseudomonas cepacia, Pseudomonas pseudomallei, Vibrio angustum,* and *Vibrio splendidus*. For one reason these were not included, please see “Appendix F: Proteins removed”, on page 373, as well as tightening of other standards - see “Species, polymorphism reduction”, on page 70, noting the problems with this run as evidence for said tightening.

[453] The harmonic mean for tree 1 (the original arrangement’s) results is very low relative to the arithmetic mean because the run’s probabilities were, for most of it, around or below the harmonic mean, then climbed to above the arithmetic mean near the end of the run. The original output files for the runs for tree arrangements 12 and 13, which would be needed to investigate the effects of other burnin settings, are unfortunately not available due to a copying error. This subset had sufficient other problems (discussed on page 241) that it was not considered worthwhile to perform the runs again.

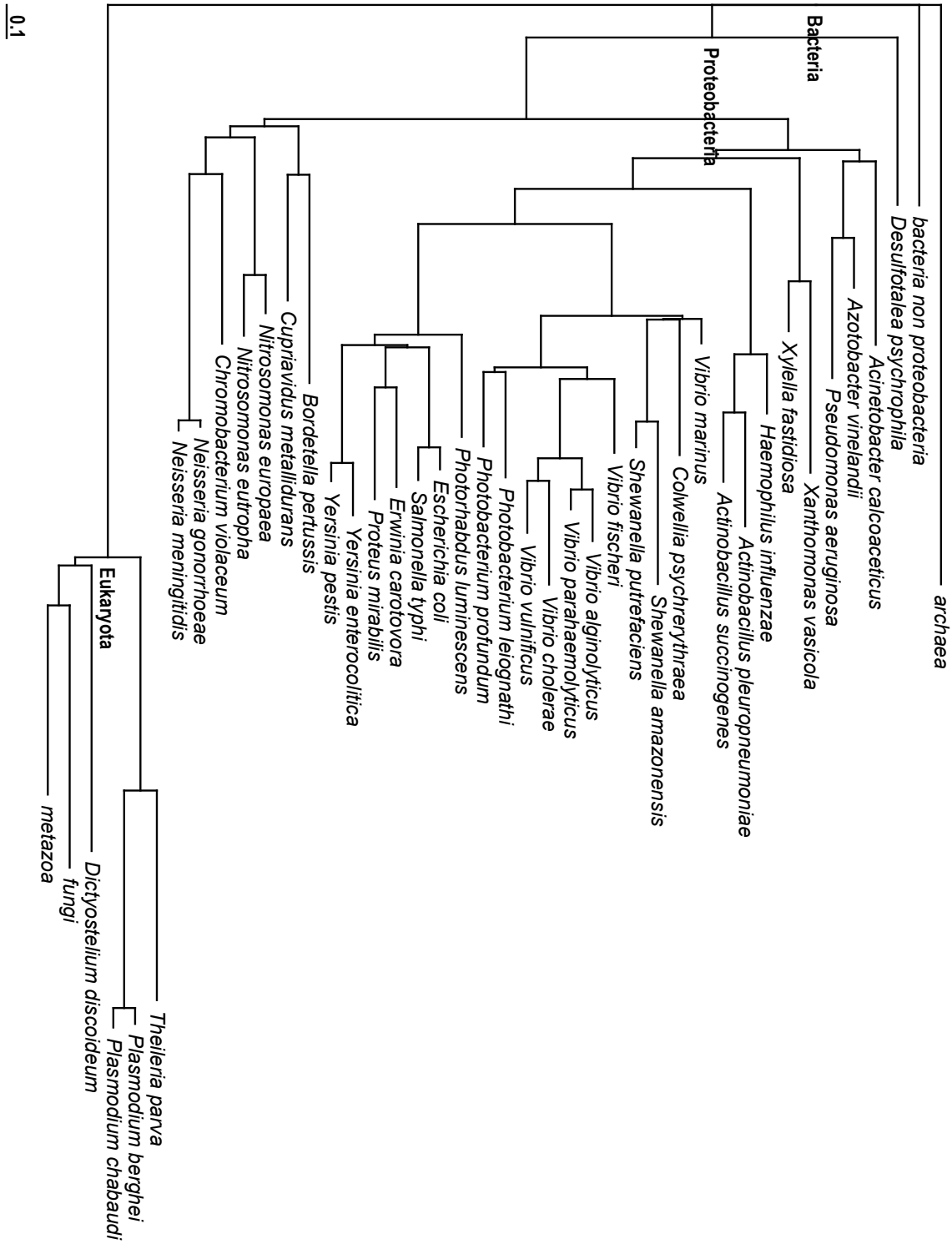


Figure 4.T.r1.s1.c.p: Round 1 subset 1 of final tree, phylogram

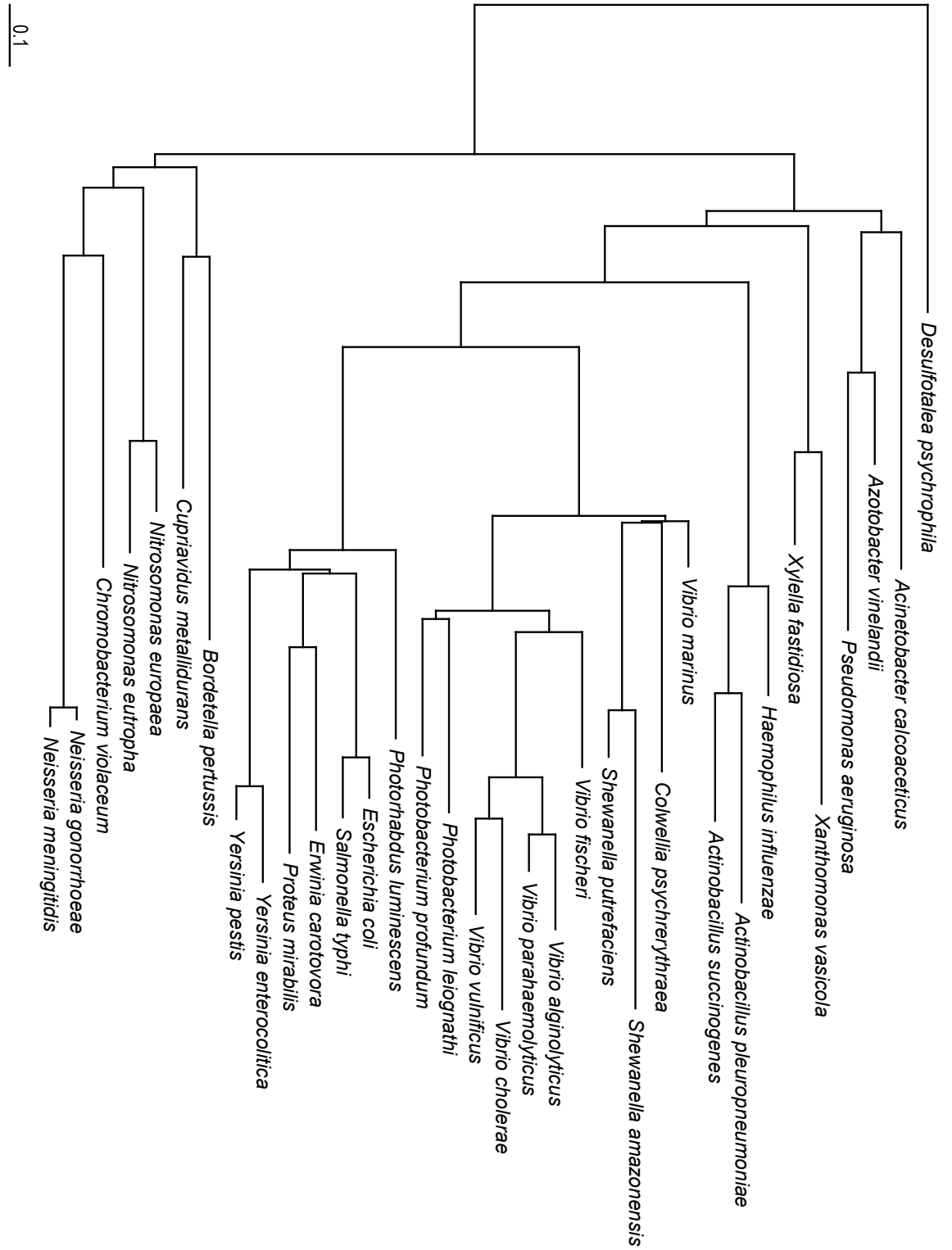


Figure 4.T.r1.s1.c.p.proteobacteria: Round 1 subset 1 of final tree, Proteobacteria only shown, phylogram

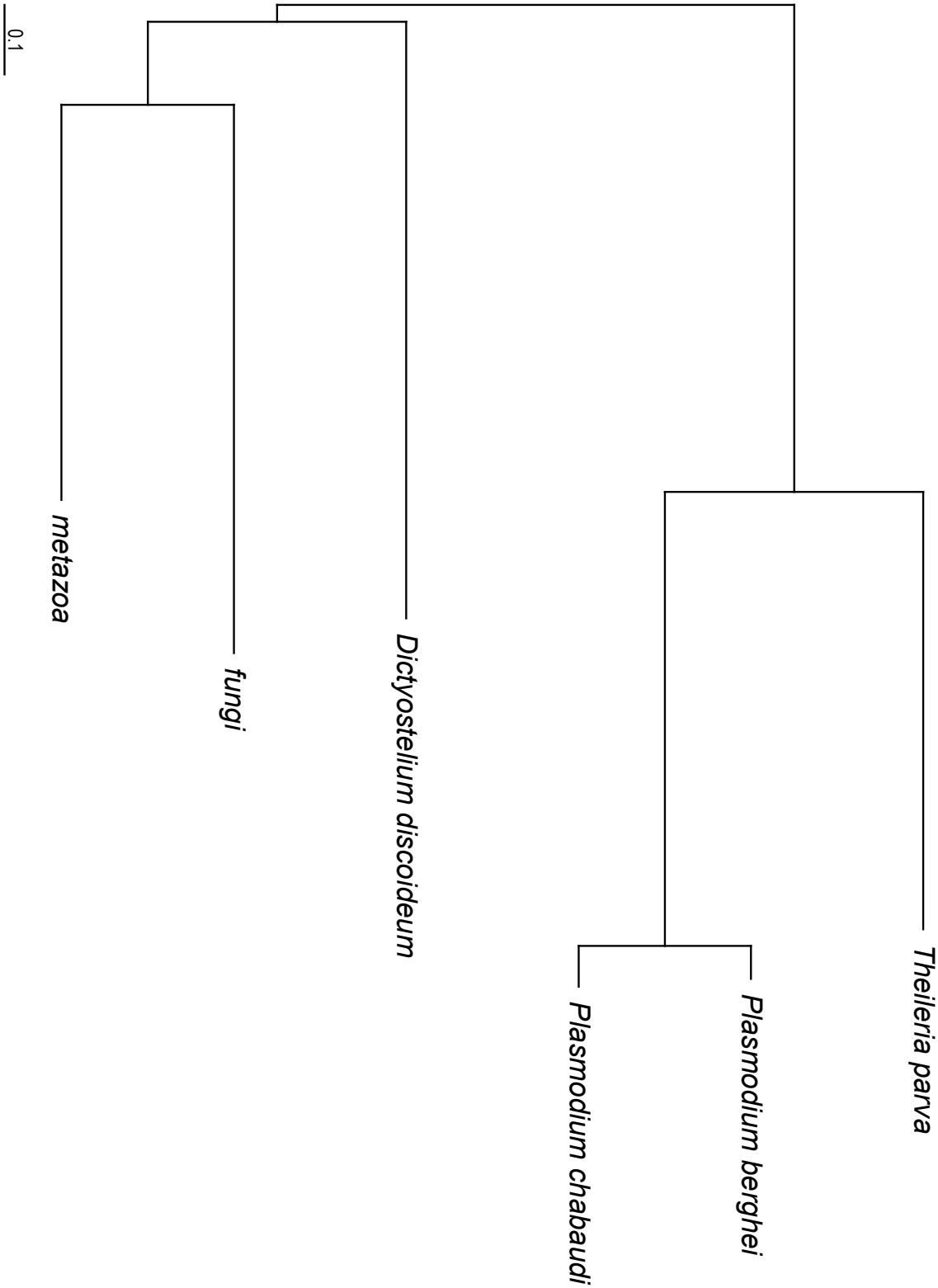


Figure 4.T.r1.s1.c.p.eukaryota: Round 1 subset 1 of final tree, Eukaryota only shown, phylogram

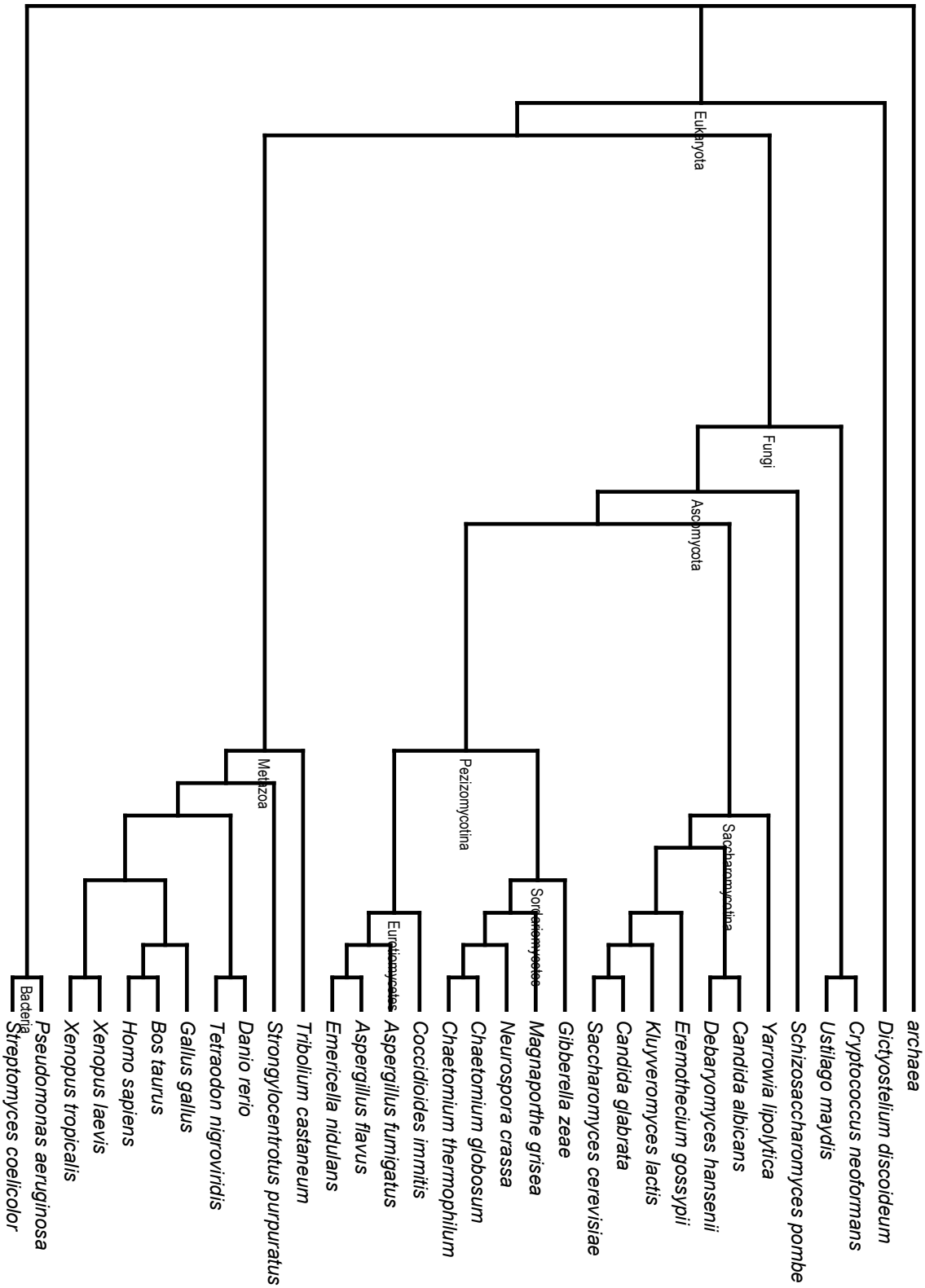


Figure 4.T.r1.s1.c.c: Round 1 subset 1 of final tree, cladogram

The following other trees with subset 1 are available (please see “Appendix L: Tree files available”, on page 394), with phylogenies as of the other subsets in round 1:

- round1.subset.1.orig.phy
- round1.subset.1.usertree.12.phy
- round1.subset.1.usertree.13.phy

Given, among other matters (e.g., see Figure 4.T.r1.s1.c.p.eukaryota, on page 239), that this subset had fungi and metazoa as groups, and the group sequence creation at the time was not using distance information (see “Further sequence processing: Group sequence creation”, on page 96)[454], the ambiguous results of this run are unsurprising. In hindsight, this subset, if used at all, should have been used solely for distance determination (for Proteobacteria).

Subset 7: Some Eukaryota (and others)

For subset 7, 10242 amino acids in 25 proteins (counting ADH1 as 1) were used for runs for 200000 generations (2000 samples); see the log probability table below for the burnins used for sumt and sump:

Phylogeny Tested	Burnin=1000		Burnin=1900	
	Arith. M.	Harmon. M.	Arith. M.	Harmon. M.
1 (original):	**-288,750.88**	**-302,927.70**	**-288,750.88**	**-290,176.11**
5:	-308,509.09	-340,918.32	-308,509.09	-309,307.12
6:	-307,015.82	-310,281.45	-307,015.82	-307,278.25
12:	-299,057.75	-307,803.46	-299,057.75	-299,605.92
13:	**-266,652.54**	**-276,861.28**	**-266,652.54**	**-267,140.09**

The species rearranged for 1 versus 5 versus 6 were *C. albicans* and *C. glabrata*. The species rearranged for 1 versus 12 versus 13 was *D. discoideum*.

[454] Of course, there would probably have been too little good distance data at this point for the distance-based algorithm to work properly. A rerun of this subset with the current algorithm (and

The final arrangement for the species[455] in this subset can be seen in Figure 4.T.r1.s7.c.p, on page 243 (a version of this with Eukaryota only shown is on page 244), and in Figure 4.T.r1.s7.c.c, on page 245. For the original (tree 1) and hypotheses 5 and 6 for subset 7, please see pages 246-251.

---

set of distances for input to it) may be of interest as a test of said algorithm.

[455] The following species are not now in the tree (due primarily to the cellulase removals - see "Appendix F: Proteins removed", on page 373), but were at the time: *Aspergillus aculeatus, Aspergillus niger, Chaetomium gracile, Fusarium oxysporum, Humicola insolens, Melanocarpus albomyces, Paecilomyces variotii, Penicillium chrysogenum, Penicillium citrinum, Penicillium funiculosum, Penicillium simplicissimum, Plectosphaerella cucumerina, Talaromyces emersonii, Thermoascus aurantiacus, Thermomyces lanuginosus, Trichoderma parceramosum, Trichoderma reesei,* and *Trichoderma viride*.

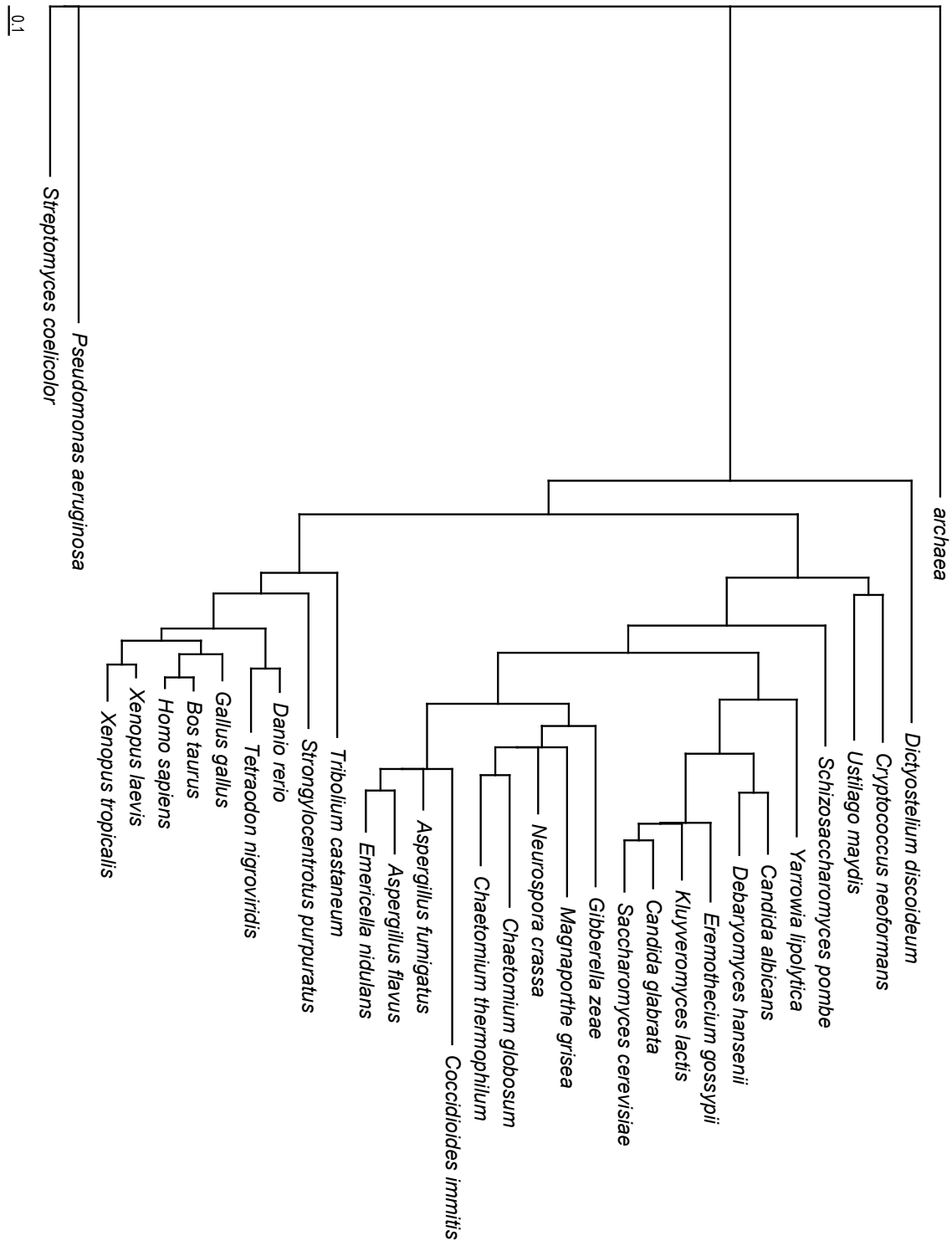


Figure 4.T.r1.s7.c.p: Round 1 subset 7 of final tree, phylogram

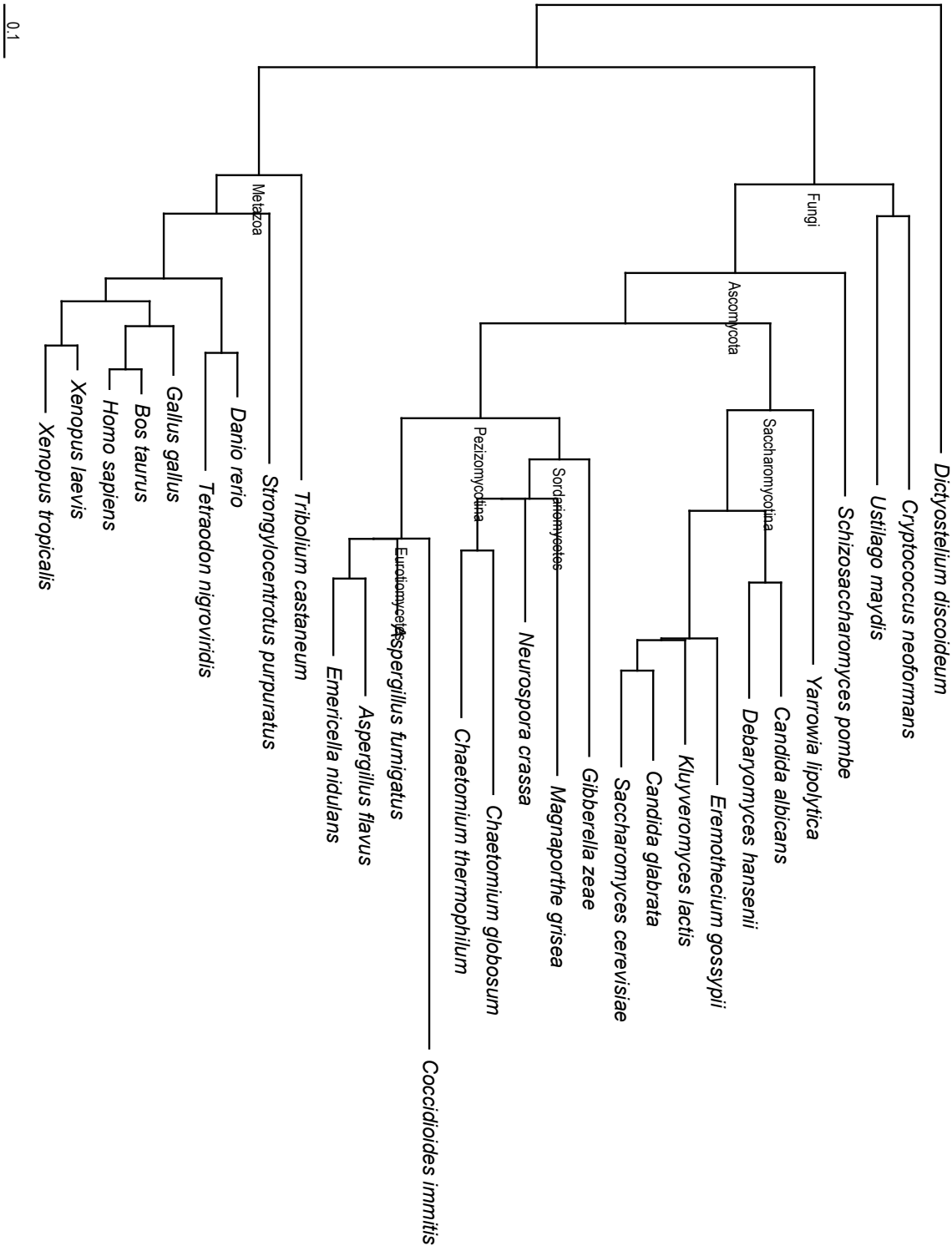


Figure 4.T.r1.s7.c.p.eukaryota: Round 1 subset 7 of final tree, Eukaryota only shown, phylogram

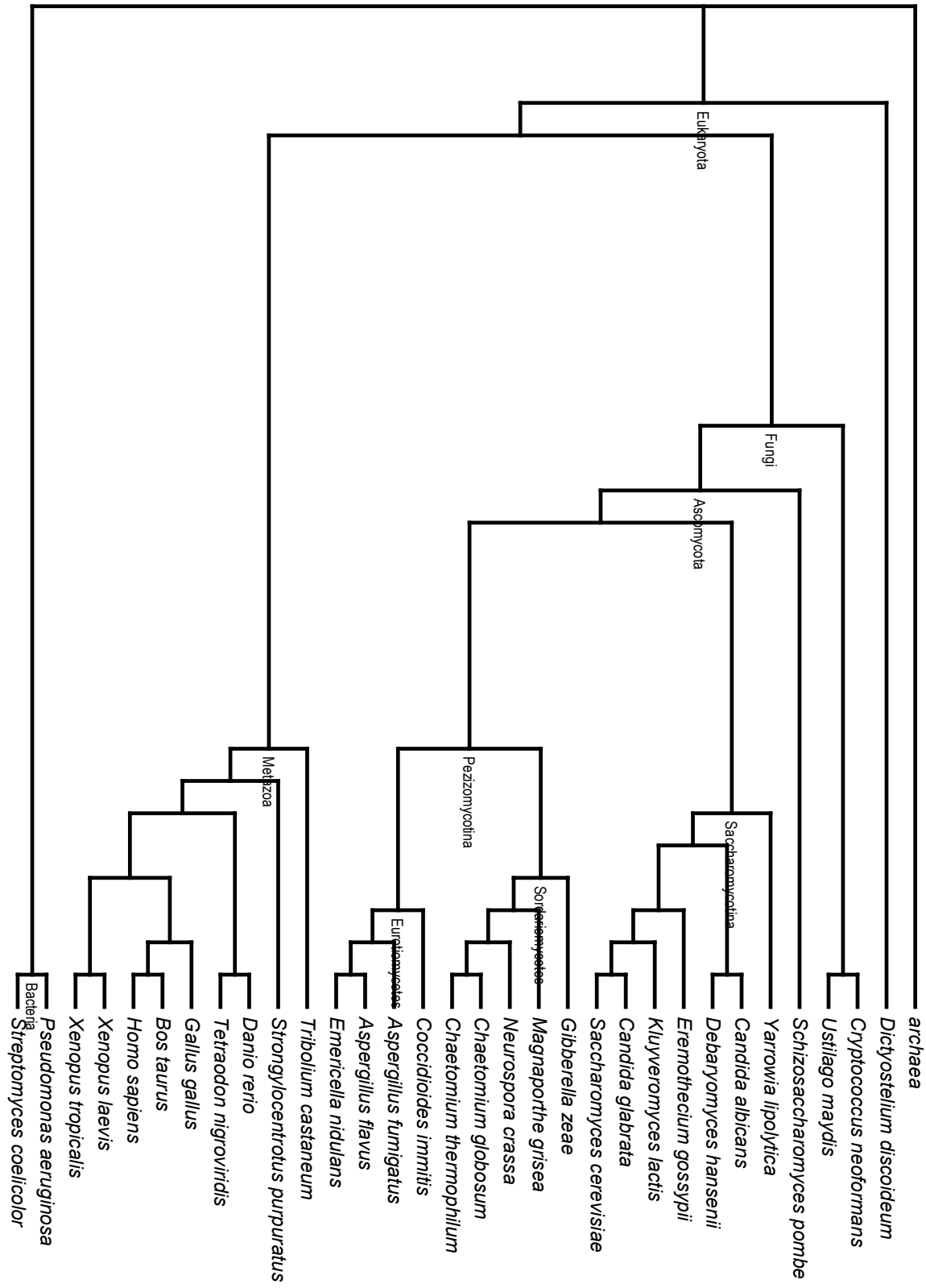


Figure 4.T.r1.s7.c.c: Round 1 subset 7 of final tree, cladogram

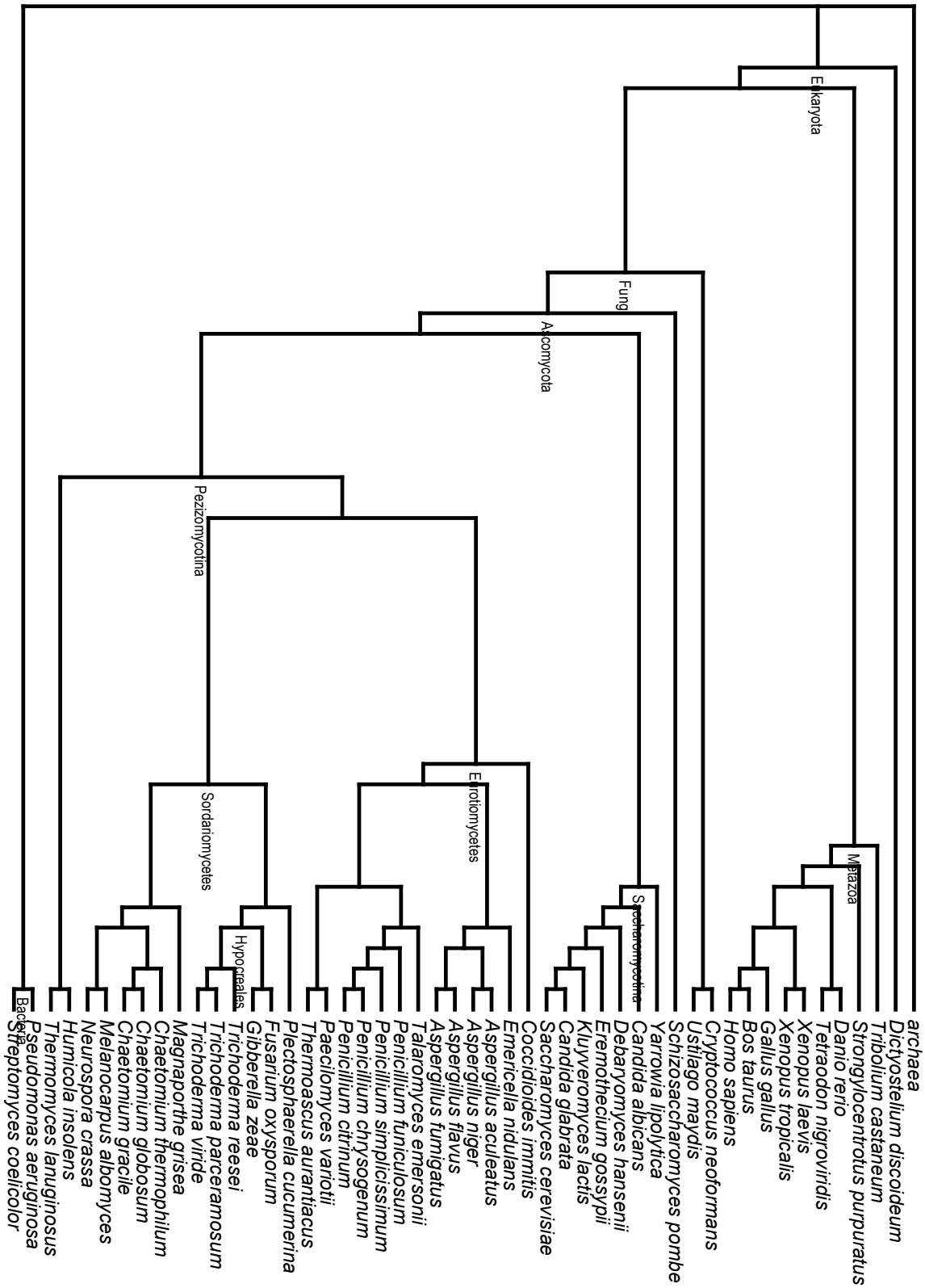


Figure 4.T.r1.s7.1: Round 1 subset 7, tree 1 (original) arrangement, cladogram

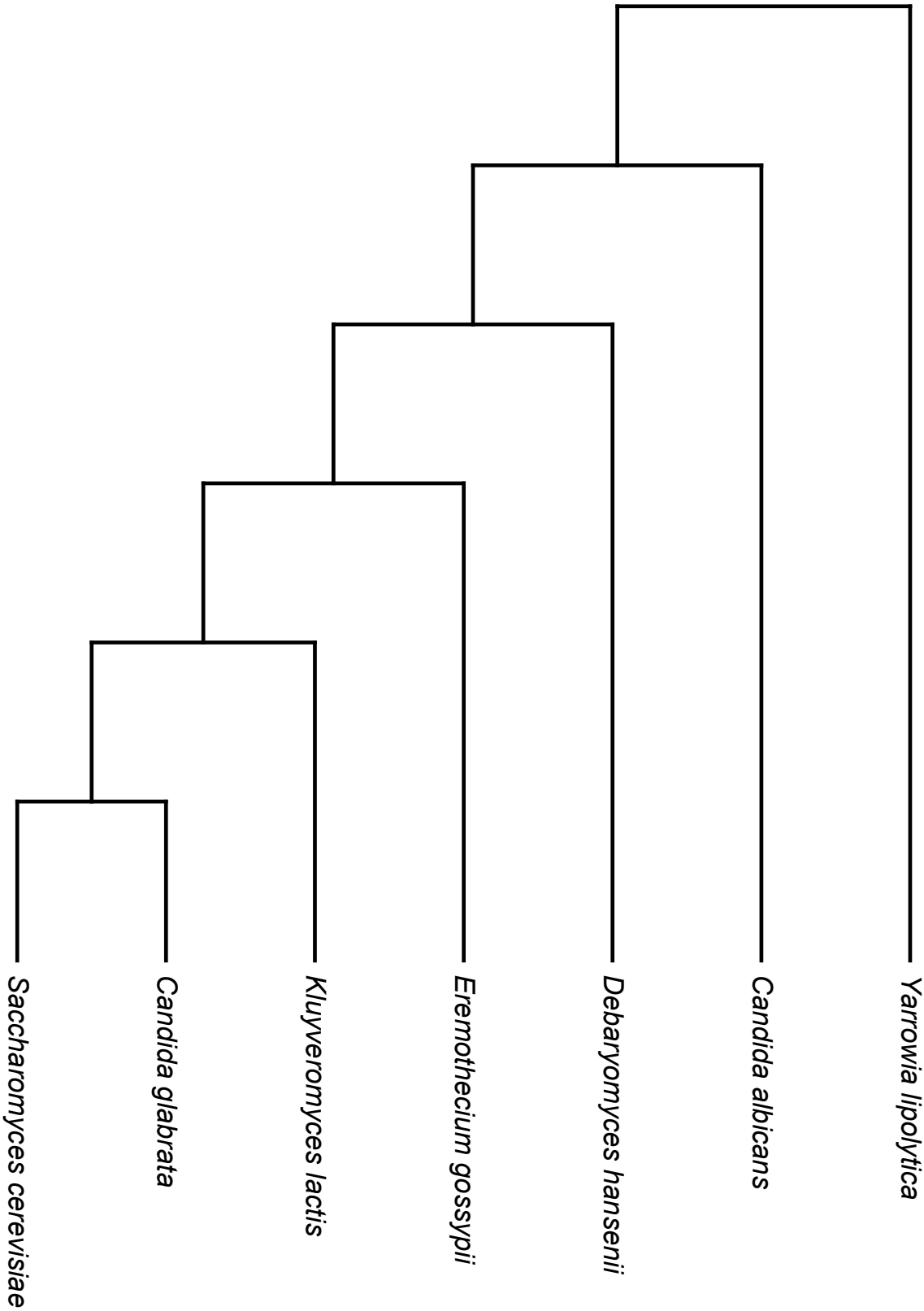


Figure 4.T.r1.s7.1.saccharomycotina: Round 1 subset 7, tree 1 (original) arrangement, Saccharomycotina only shown, cladogram

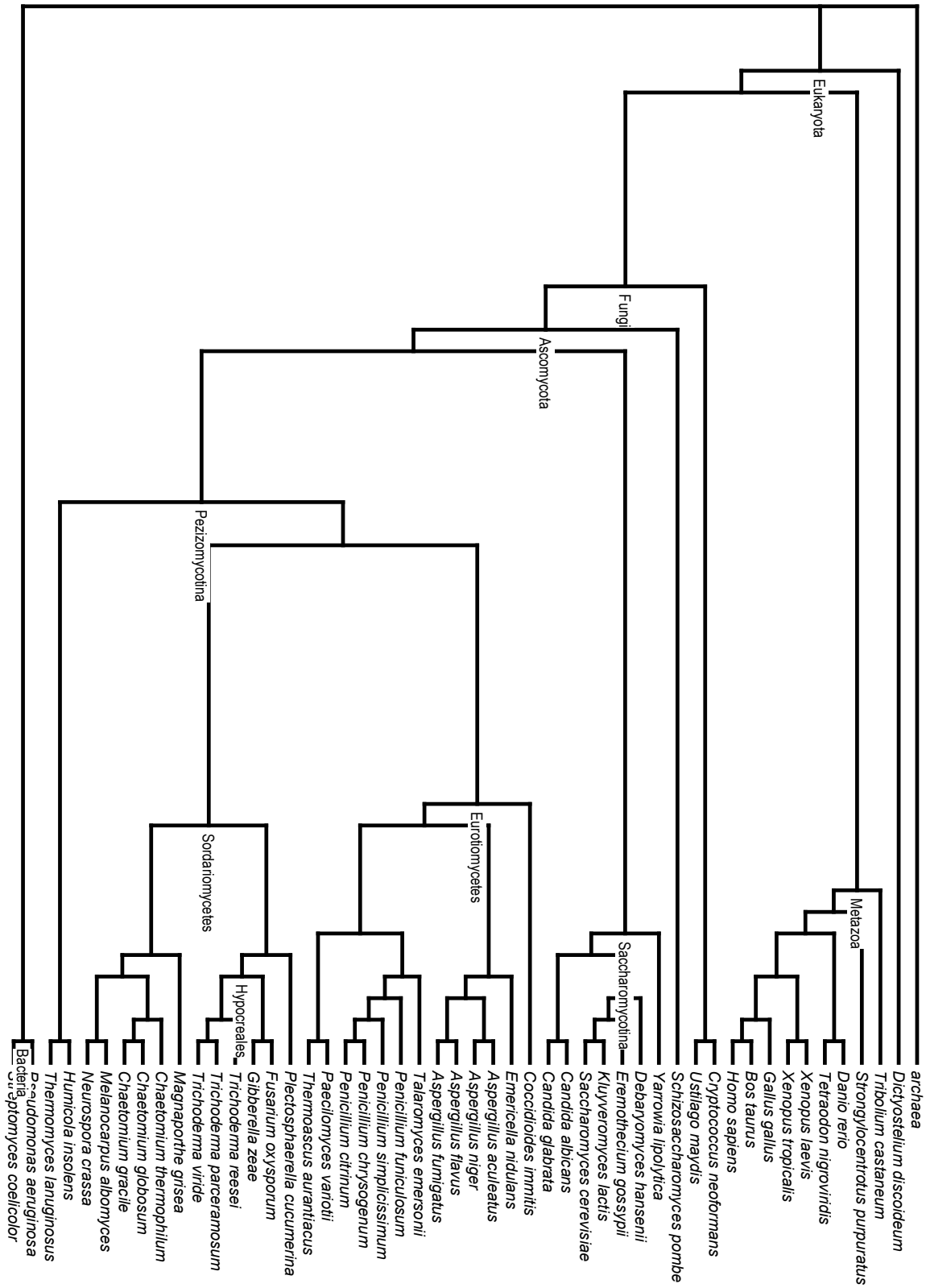


Figure 4.T.r1.s7.5: Round 1 subset 7, tree 5 arrangement, cladogram

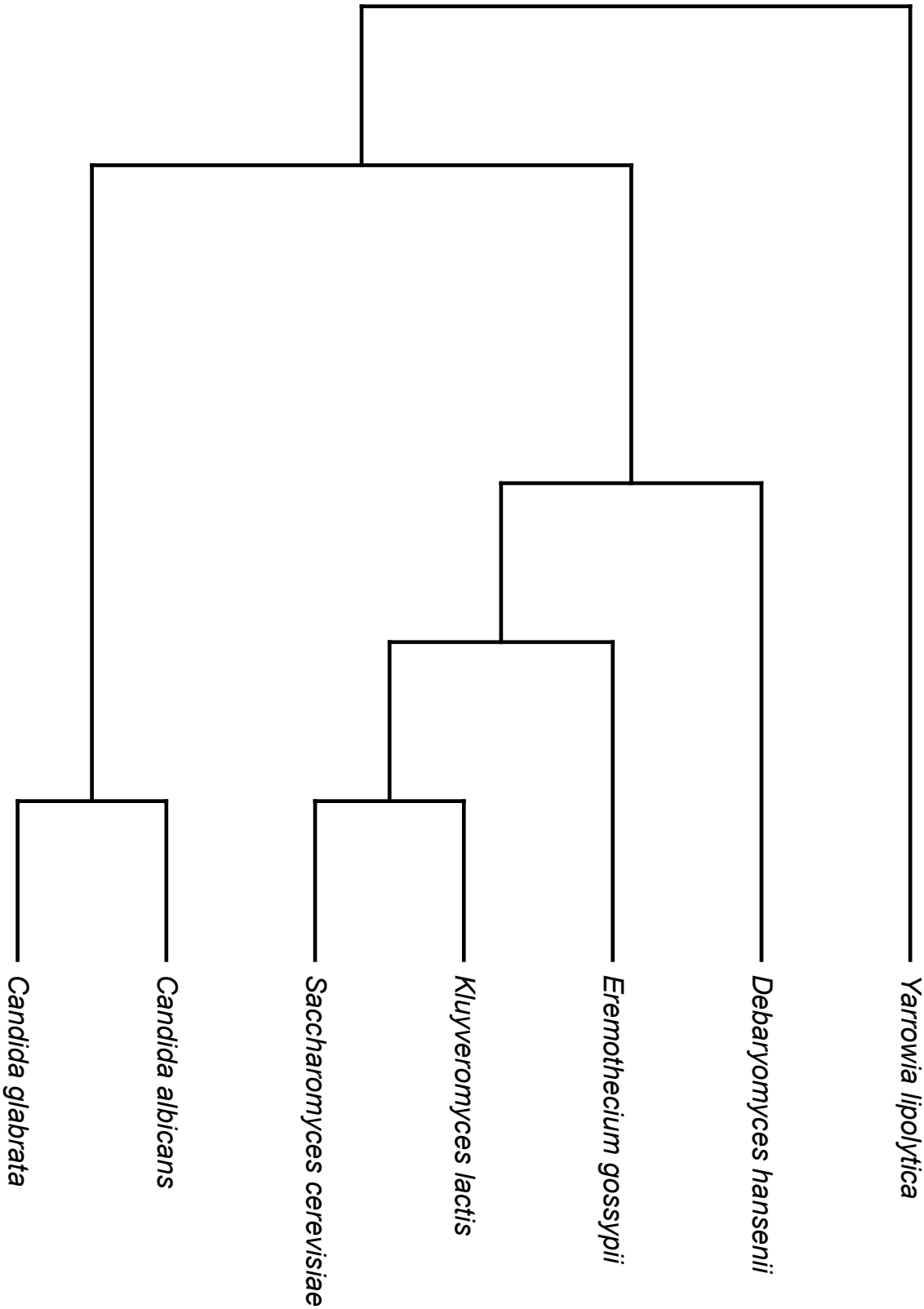


Figure 4.T.r1.s7.5.saccharomycotina: Round 1 subset 7, tree 5 arrangement, Saccharomycotina only shown, cladogram

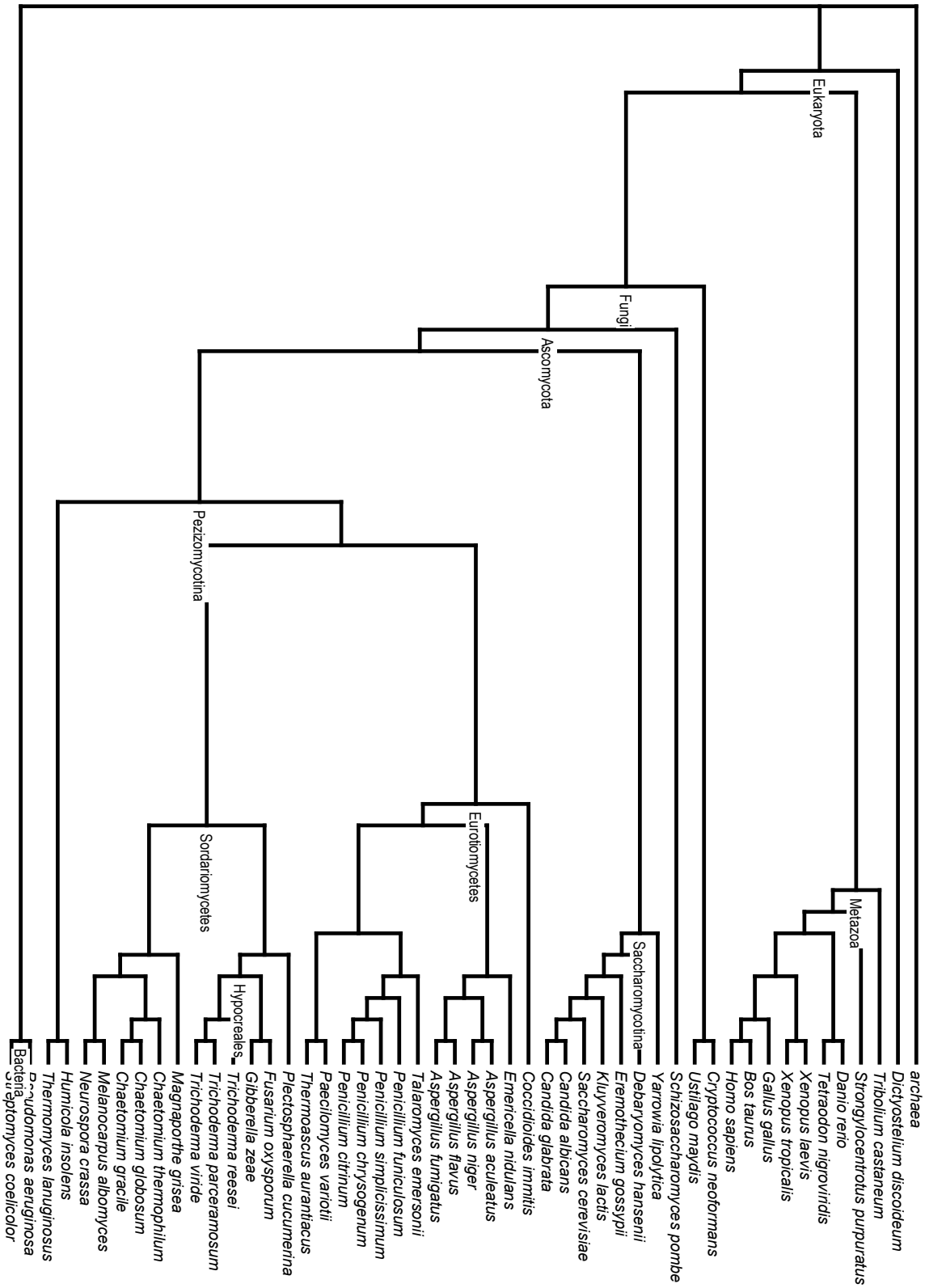


Figure 4.T.r1.s7.6: Round 1 subset 7, tree 6 arrangement, cladogram

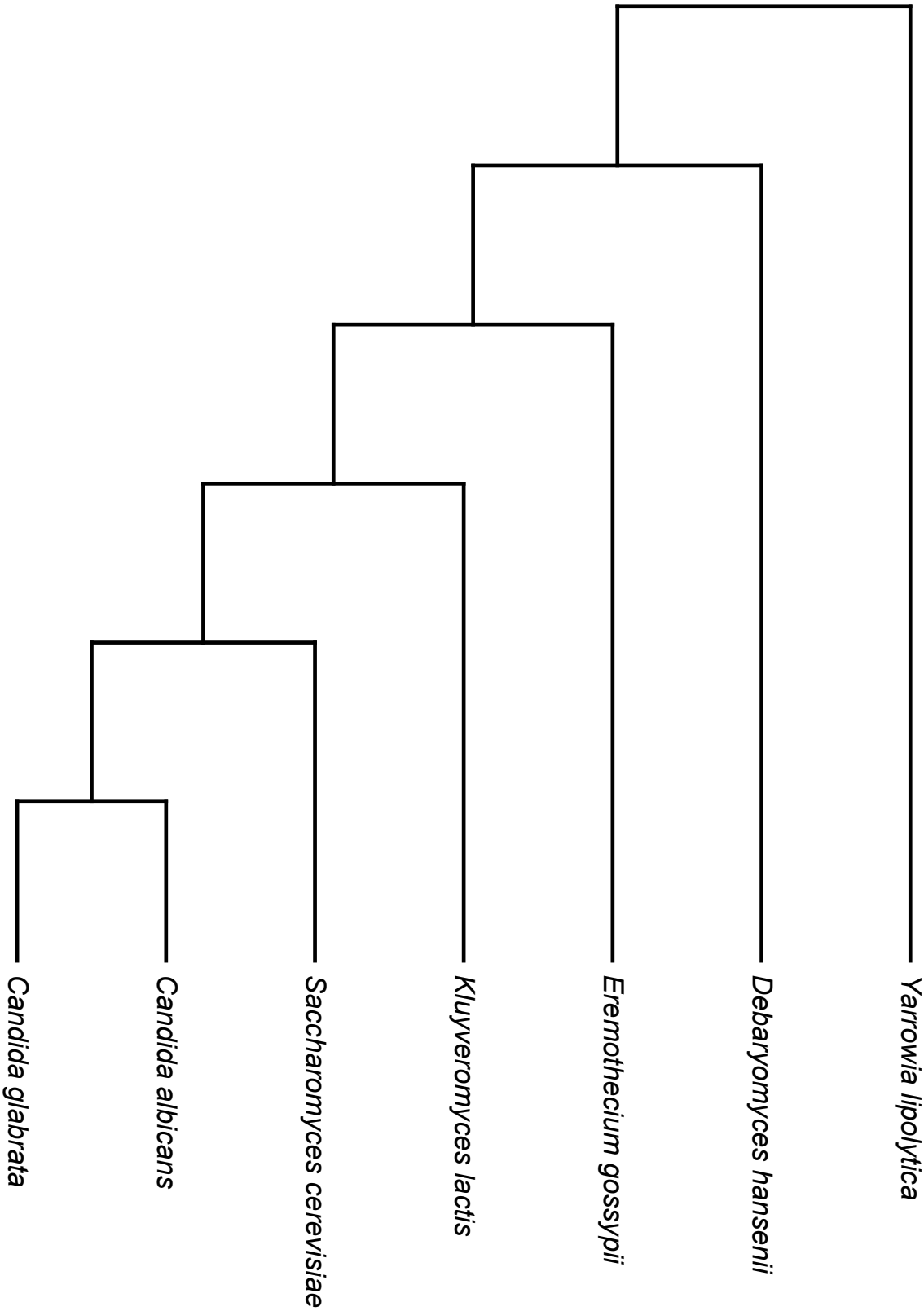


Figure 4.T.r1.s7.6.saccharomycotina: Round 1 subset 7, tree 6 arrangement, Saccharomycotina only shown, cladogram

For other trees with subset 7, round 1, see (via "Appendix L: Tree files available", on page 394):

- round1.subset.7.usertree.12.phy
- round1.subset.7.usertree.13.phy

It was originally thought from the combination of the above for hypotheses 12 and 13 that *D. discoideum* might be closer to fungi/metazoa than *E. histolytica,* given that subsets with only the first present (6, 1, and 7) gave a result different from that of when both were present (subset 2). However, subsequent results (e.g., "Tree search with Eukaryota (subset)", on page 300) indicated otherwise[456].

Subset 3: Some Eukaryota, Bacteria

Subset 3 had 10298 amino acids, from 25 proteins (counting ADH1 as 1), and was used for runs with 200000 generations (2000 samples). Please see the log probability table below for the burnins used for sump and sumt:

Phylogeny Tested	Burnin=1000		Burnin=1750	
	Arith. M.	Harmon. M.	Arith. M.	Harmon. M.
1 (orig):	**-147,734.10**	-190,507.00	**-147,734.10**	**-156,532.15**

[456] Admittedly, long-branch attraction between these two species is possible. For future work, we recommend the use of one or more of:

- more proteins (e.g., actin, as previously mentioned - see, among others, footnote 234, on page 113)
- more (putatively) nearby species
- the covarion option (see footnote 200 under "MrBayes code alterations", on page 99) if it can be made to work with (structurally-aligned, ideally) rRNA (or tRNA, etc.), including considerations of stem/loop structure, as opposed to protein sequences
- tree searches run with less constraints (see footnote 468 under "Tree search with Eukaryota (subset)", on page 303)
- other methods of overcoming long branch attraction and similar problems (see "Future work", on page 334)

Note, however, that these species were considered less important for the current work, despite their interesting putative phylogenetic position (near the root of fungi/metazoa), due to their lack of (usable/locatable) DHFR sequences.

Phylogeny Tested	Burnin=1000		Burnin=1750	
	Arith. M.	**Harmon. M.**	**Arith. M.**	**Harmon. M.**
5:	-167,300.85	-238,909.50	-167,300.85	-167,534.20
6:	-171,450.50	**-178,992.38**	-171,450.50	-176,685.01

The species moved between arrangements 1, 5, and 6 were all members of the current "*Candida*" genus. On pages 254-262 are the trees for subset 3.

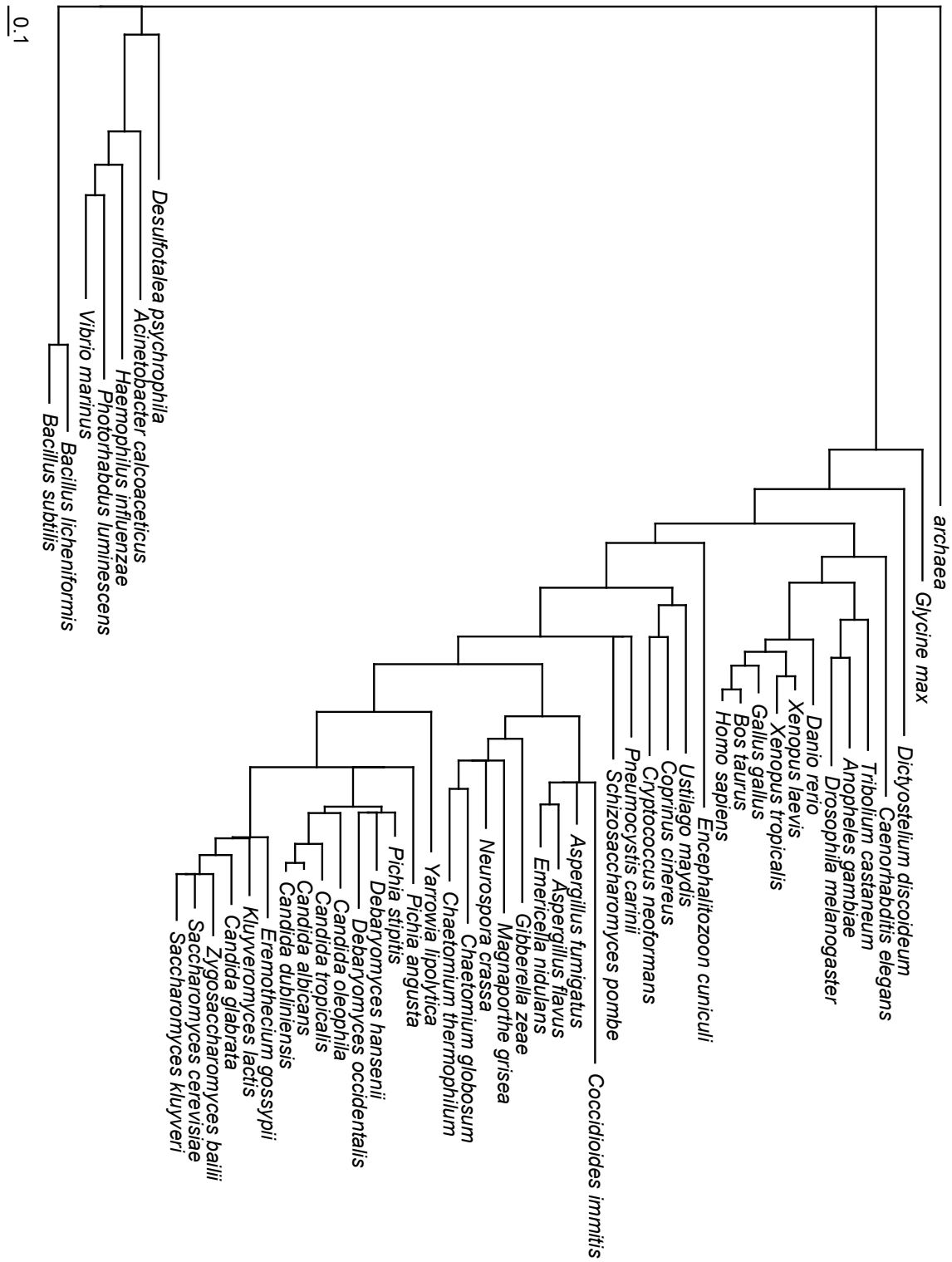


Figure 4.T.r1.s3.c.p: Round 1 subset 3 of final tree, phylogram

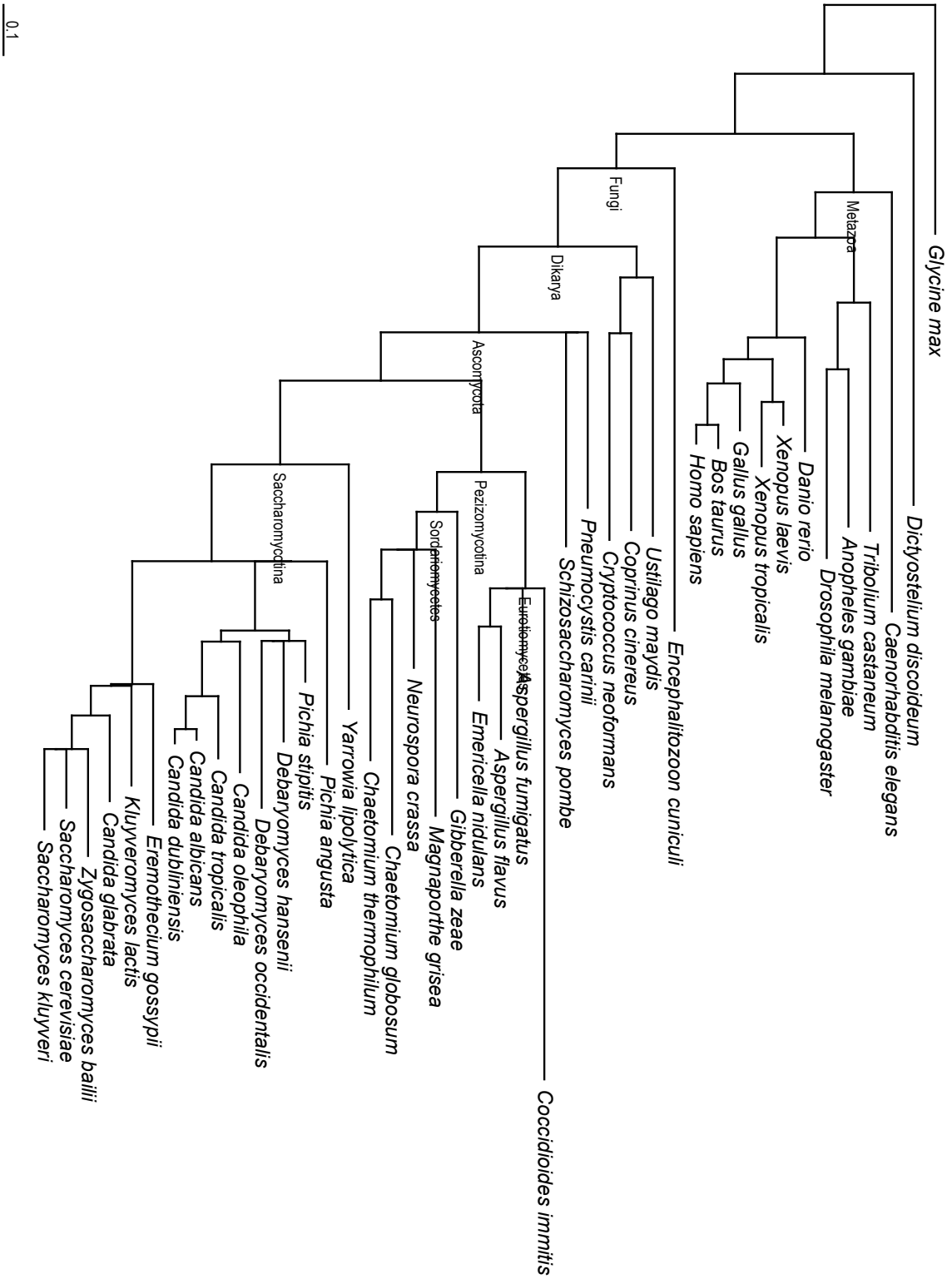


Figure 4.T.r1.s3.c.p.eukaryota: Round 1 subset 3 of final tree, Eukaryota only shown, phylogram

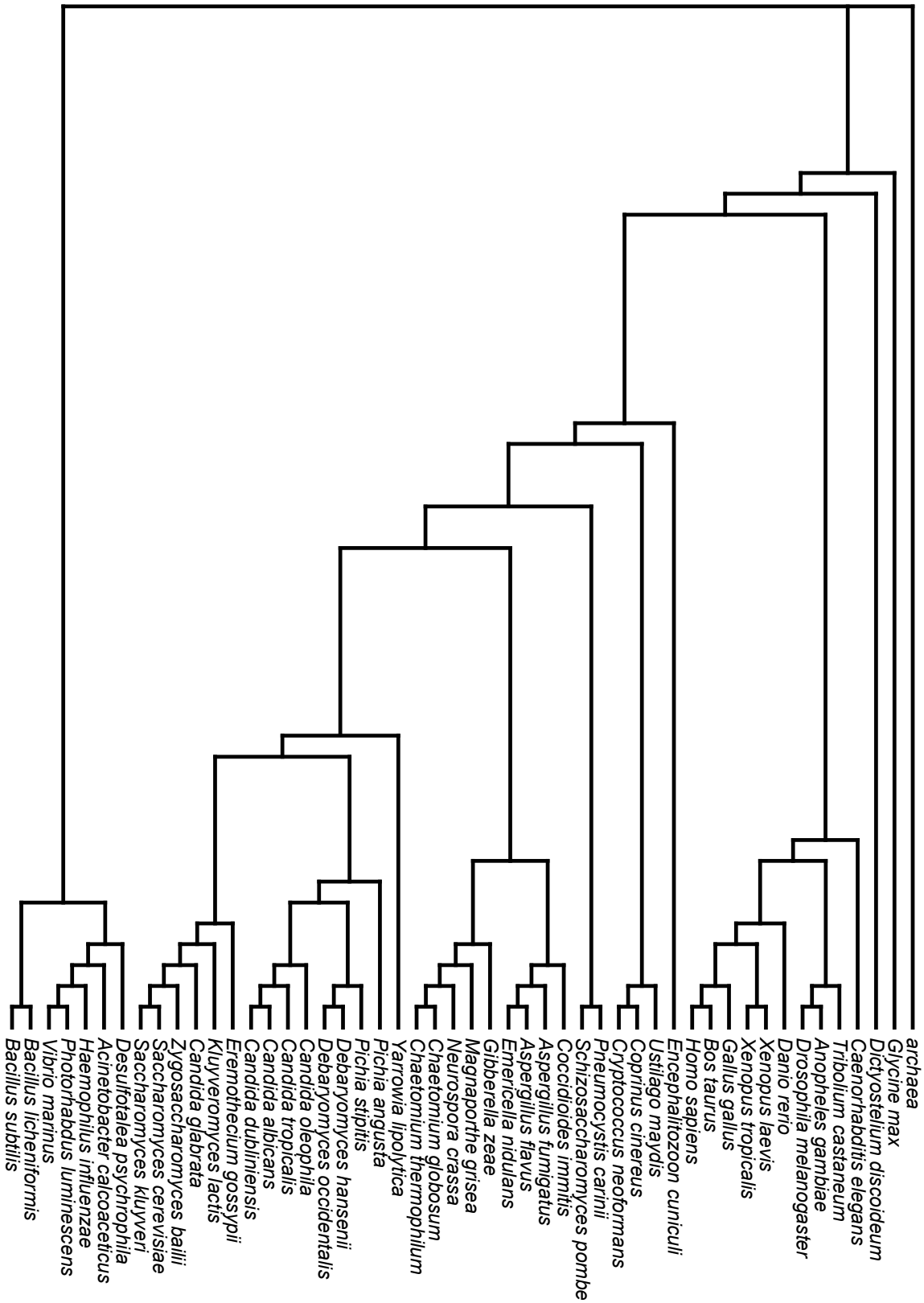


Figure 4.T.r1.s3.c.c: Round 1 subset 3 of final tree, cladogram

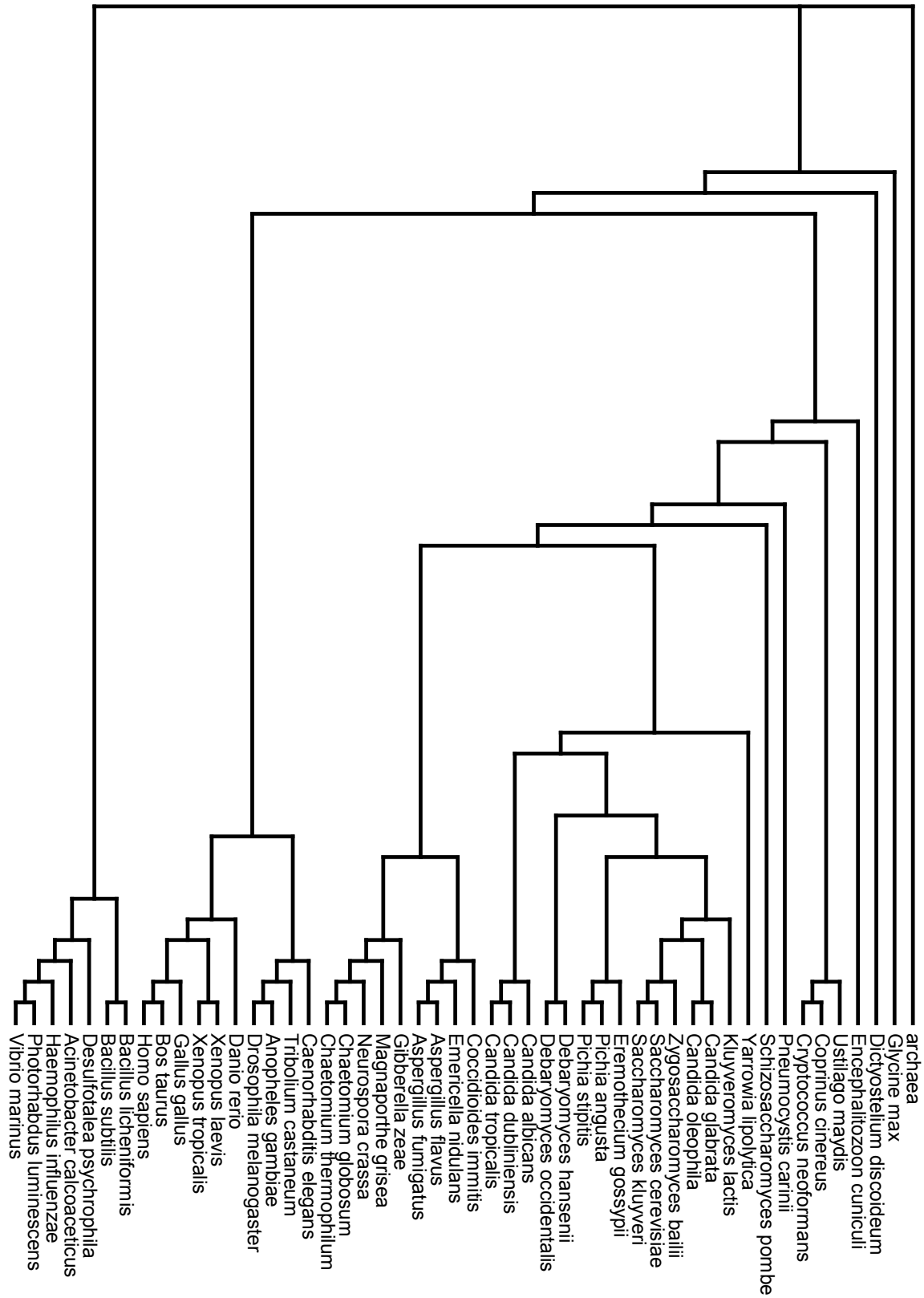


Figure 4.T.r1.s3.1: Round 1 subset 3, tree 1 (original) arrangement, cladogram

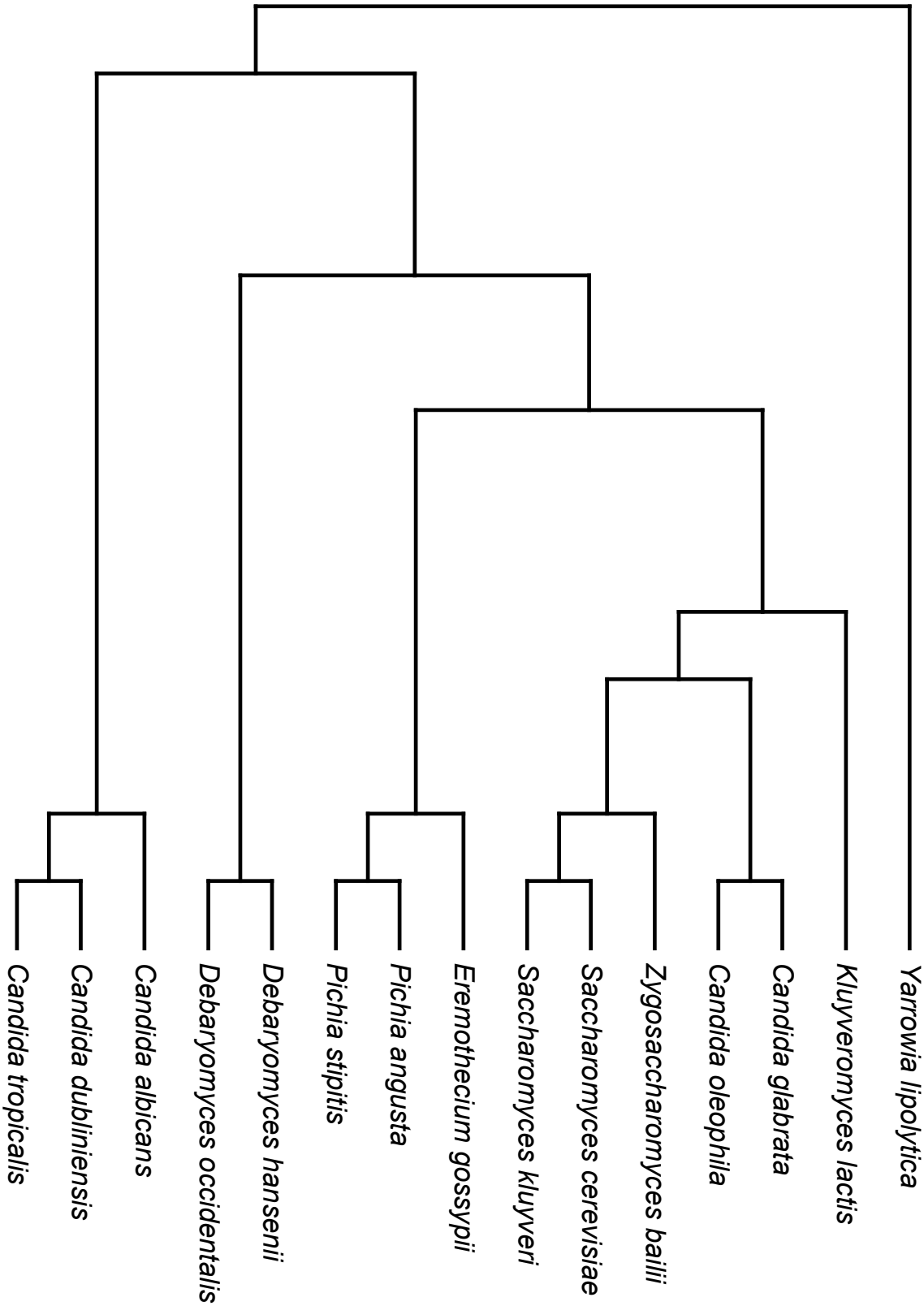


Figure 4.T.r1.s3.1.saccharomycotina: Round 1 subset 3, tree 1 (original) arrangement, Saccharomycotina only shown, cladogram

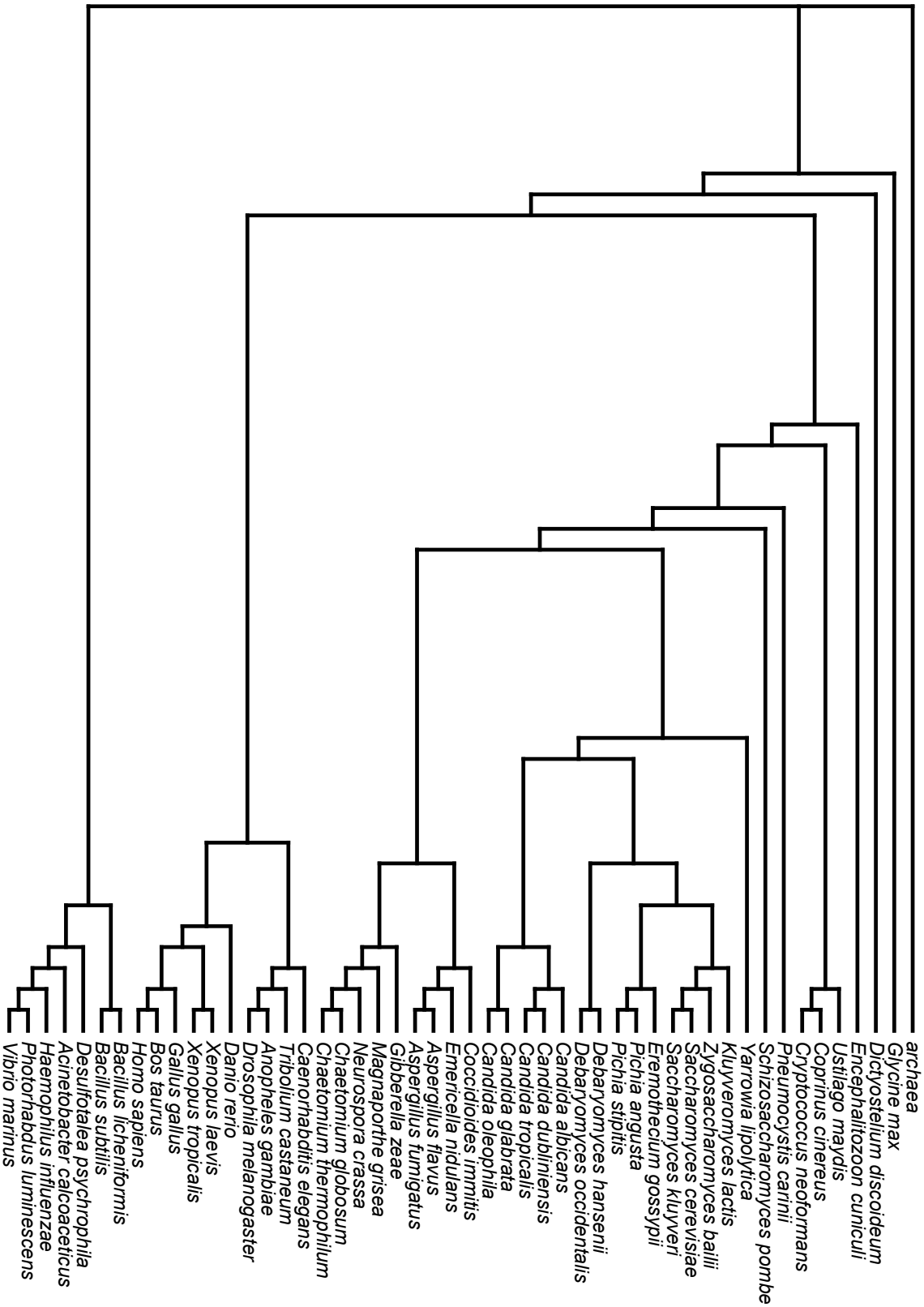


Figure 4.T.r1.s3.5: Round 1 subset 3, tree 5 arrangement, cladogram

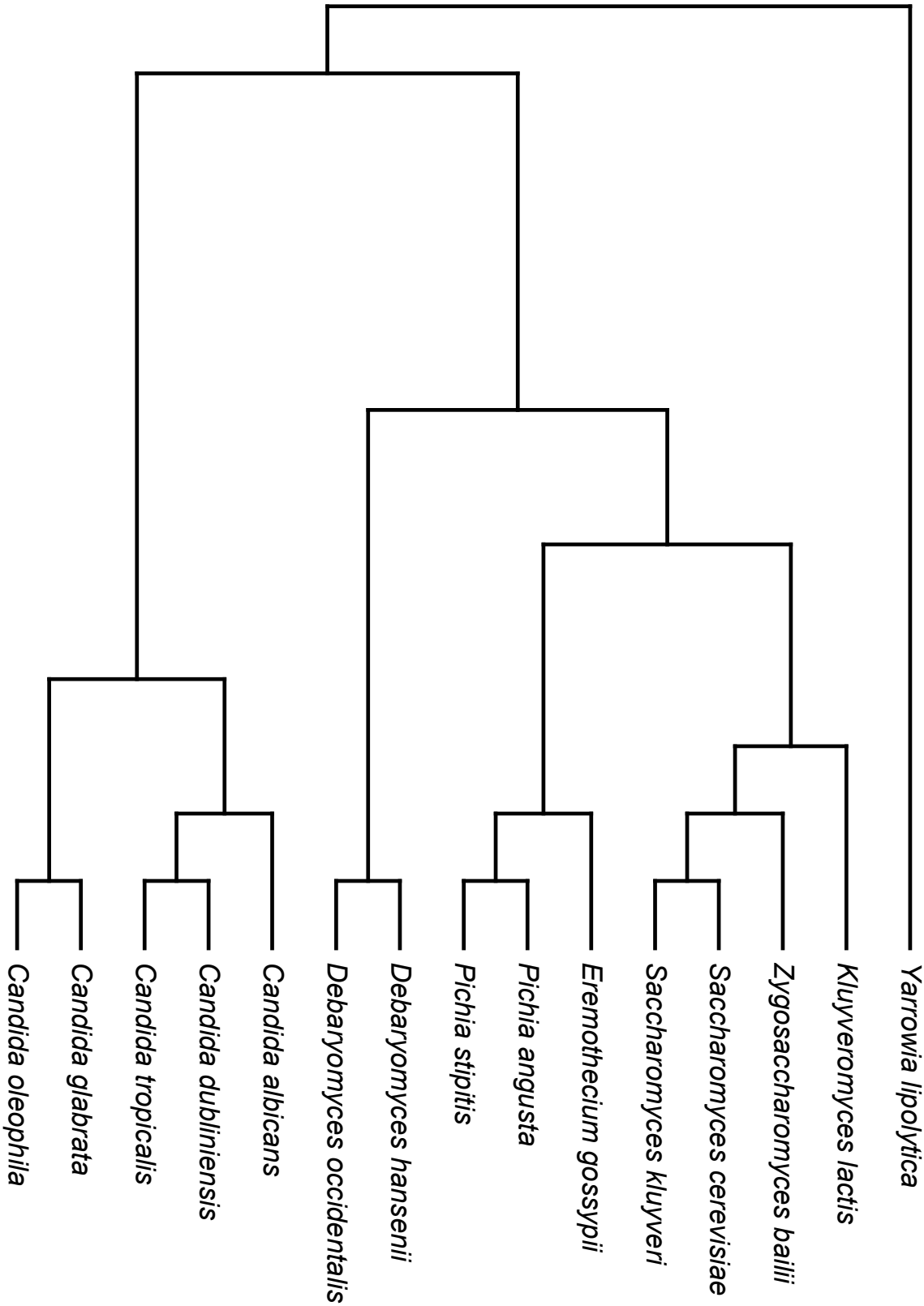


Figure 4.T.r1.s3.5.saccharomycotina: Round 1 subset 3, tree 5 arrangement, Saccharomycotina only shown, cladogram

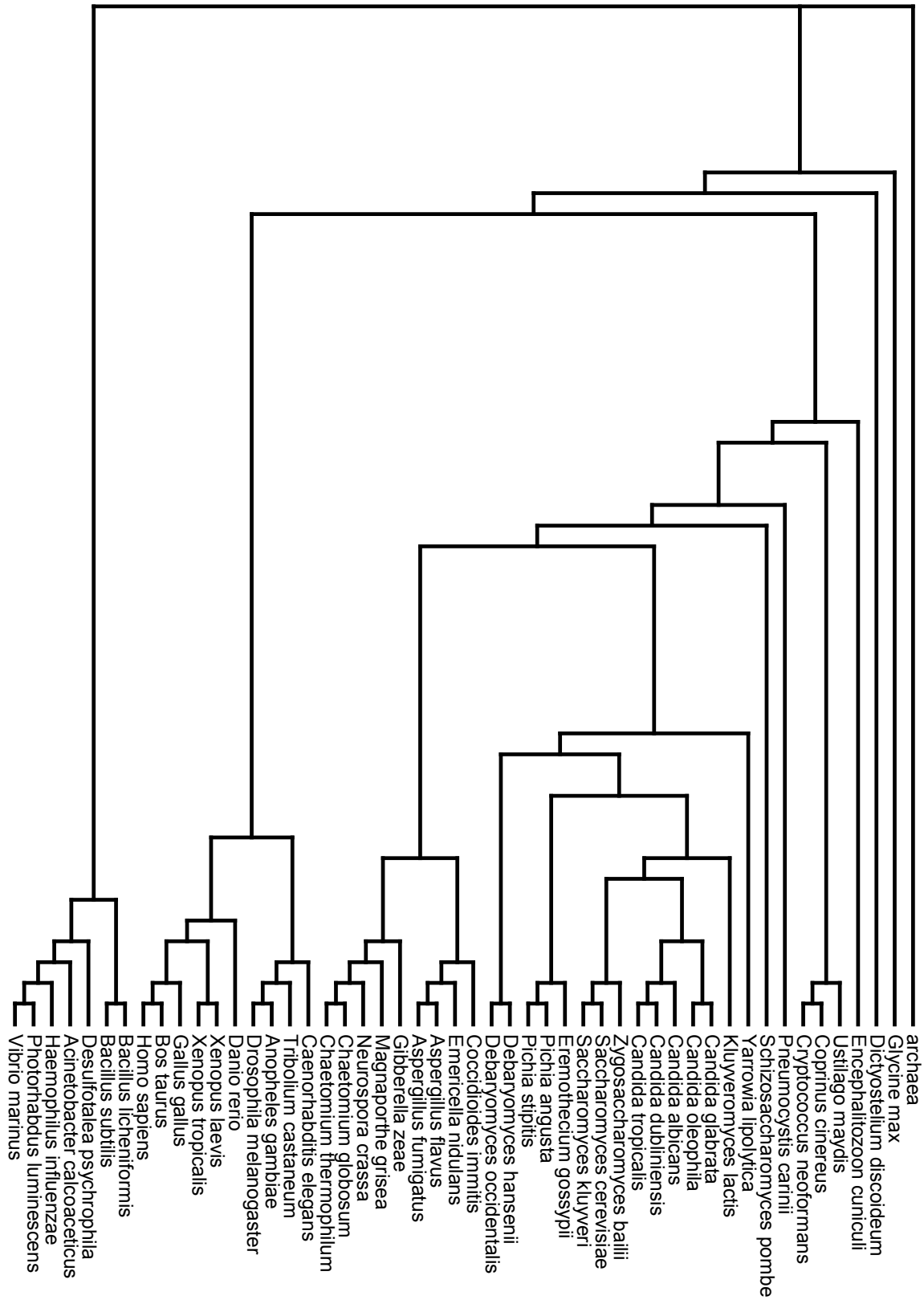


Figure 4.T.r1.s3.6: Round 1 subset 3, tree 6 arrangement, cladogram

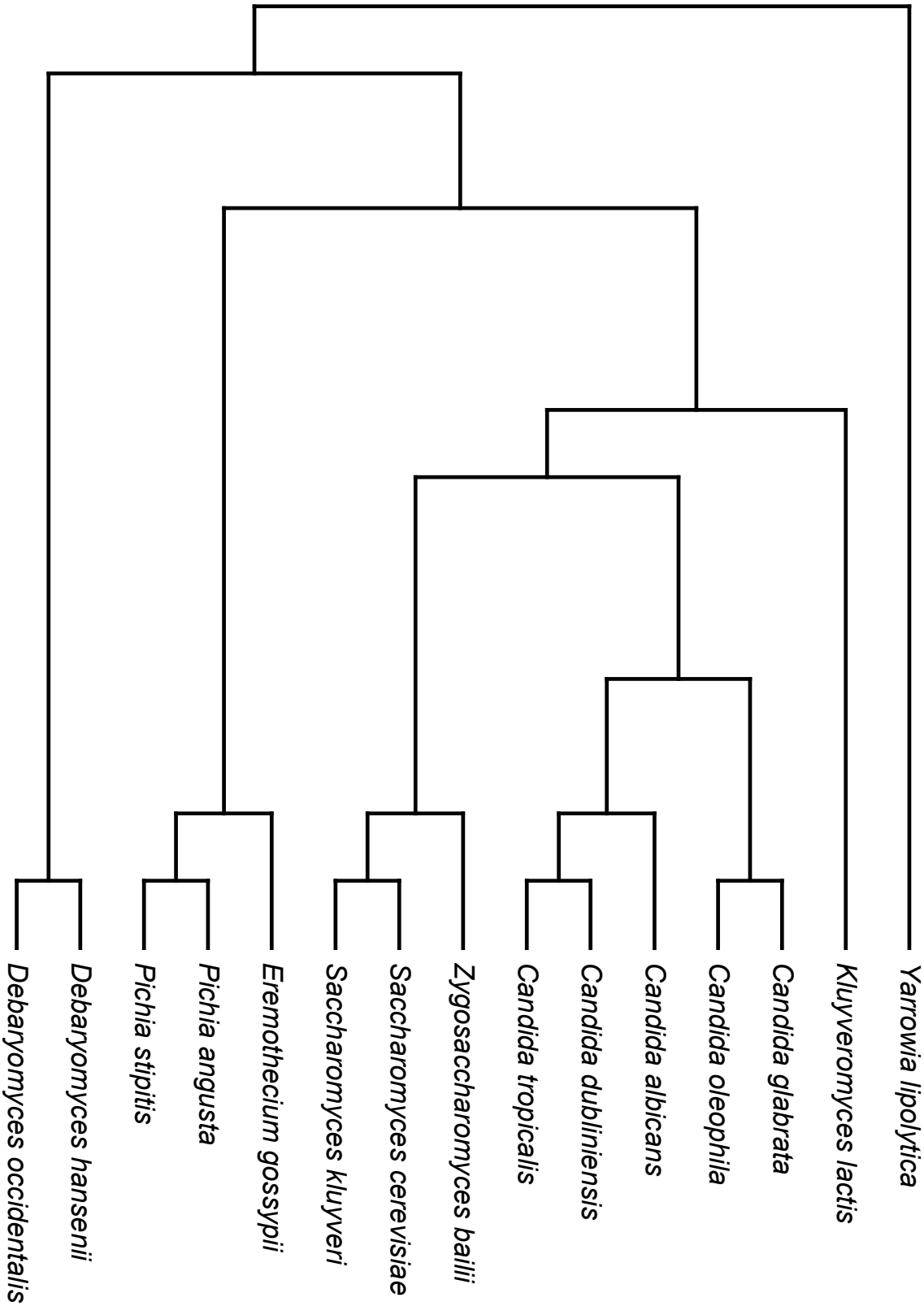


Figure 4.T.r1.s3.6.saccharomycotina: Round 1 subset 3, tree 6 arrangement, Saccharomycotina only shown, cladogram

The overall conclusion from the above (phylogeny 1 better than 5 or 6) would be that it is probable that the current *Candida* genus is not a clade, with a slight possibility that it is one but more closely related to *S. cerevisiae* than previously thought. Other subsets with information on the subject include subset 7 (see “Subset 7: Some Eukaryota (and others)”, on page 241); it appears that subset 7 “agrees” with subset 3 that the current *Candida* genus is not a clade.

Subset 4

An additional subset, 4, was also tested, but this testing more points out a problem with tree rearrangements than serves as anything useful. The testing in question was of whether Kinetoplastida (*Leishmania* and *Trypanosoma*, in our dataset) and Viridiplantae should be branching together, as opposed to Alveolata (e.g., *Plasmodium*) branching with Kinetoplastida branching off together after Viridiplantae branched off. The results appeared to say yes, but the final tree used contradicts both of these, with Viridiplantae branching off from other eukaryota - in our dataset - prior to other branching (see “Tree search with Non-Fungi/Metazoa Eukaryota”, on page 313). This may point to the utility of doing at least two different subsets for any given phylogenetic question to be asked by rearrangements, as a variety of bootstrapping (see under “2. Phylogenetics - Ancestral Sequence Prediction”, footnote 20, on page 12) - if datasets contradict each other, then this is an area to be investigated more closely.

Summary of first round results

The below table is a summary of the tree rearrangement (hypothesis) results (see "First round of tree rearrangements", on page 203) from each subset in this round, with boldface indicating the stronger of two results when applicable:

Subset	1 vs. 5 vs. 6	1 vs. 12 vs. 13	1 vs. 15	1 vs. 2|3|4
2	N/A	**1** (e+d)	**1**	N/A
5	N/A	N/A	N/A	**4**
6	N/A	12 or **13** (d)	N/A	**2|3|4**
1	N/A	1 or 12 (d)	N/A	N/A
7	**1**	**13** (d)	N/A	N/A
3	**1** or 6	N/A	N/A	N/A

In the above, "(e+d)" indicates that both *E. histolytica* and *D. discoideum* were present in the subset, while a "(d)" indicates only *D. discoideum* was present. As a summary of the conclusions:

- 1 versus 5 versus 6 was testing the positions of *C. albicans* and *C. glabrata* relative to each other and to *S. cerevisiae*. The conclusion (tree 1) was that *C. albicans* was further away from *S. cerevisiae*, whereas *C. glabrata* and *S. cerevisiae* were close together.
- 1 versus 12 versus 13 was testing the positions of *D. discoideum* and *E. histolytica* vis-à-vis Fungi and Metazoa. The conclusion was that either (1) they branched off prior to Fungi and Metazoa, or (13) at least *D. discoideum* branched off Metazoa after Fungi branched off.
- 1 versus 15: This concluded that the existing arrangement of Bacterial groups (1) was better than the alternative tried (15).
- 1 versus 2|3|4: This concluded that the "classical" arrangement of Acoelomata, Pseudocoelomata, and Coelomata as three clades was correct (2|3|4), but

that Pseudocoelomata branched off before Acoelomata and Coelomata (4), contrary to the "classical" arrangement.

Second round of tree rearrangements

The possible tree rearrangements (hypotheses about organismal descent) tested by each phylogenetic comparison done for round 2 are as follows[457]:

- 1 versus 2 versus 3 - This is a comparison of:
  - *D. discoideum* at a position closer to fungi/metazoa than *E. histolytica*, but outside of the fungi/metazoa grouping (1);
  - *D. discoideum* being closer to Metazoa than Fungi (2, noting that 2 has other rearrangements as explained in footnote 457);
  - *D. discoideum* and *E. histolytica* together (3 - note the error discussed in footnote 457).
- 1 versus 2, 5 - This is a comparison of either *Debaryomyces hansenii* located with other species with a CUG serine (1), or back in its earlier position closer to *S. cerevisiae* (2 and 5).

[457] Tree 2 was actually the initial tree produced from examining the results from the first round. This was altered by changes in the positions of:
- *D. discoideum* (using a more conservative rearrangement from the former one, due to prior data (Baldauf & Doolittle 1997));
- *Debaryomyces hansenii*, *Pichia styptis*, and *Candida oleophila* (due to the discovery of research indicating their sharing with *C. albican*s of the CUG codon coding for serine (Fitzpatrick *et al.* 2006; Sugita & Nakase 1999a, 1999b))

The result was treated as tree 1, but tree 2 was kept for comparisons given the number of changes, and some of these were tested independently as a part of the rearrangements. These tests had been though to include of the first (for *D. discoideum*), via rearrangement 3, but due to an error when editing this tree (see "Future work", on page 334), it was actually changed in the opposite direction (see above).

- 1 versus 9 versus 10 - This rearrangement set is to try to determine the position of Cetartiodactyla[458], the options tried being:
  - 1: Cetartiodactyla branching first, then Carnivora (e.g., *Canis lupus*) then Primates and Rodentia together;
  - 9: Carnivora branching first, then Cetartiodactyla, then Primates and Rodentia
  - 10: Carnivora branching first, then Rodentia, then Cetartiodactyla and Primates
- 1 versus 11 versus 12 - This rearrangement set is to try to determine the positions of Viridiplantae+Kinetoplastida and Alveolata vis-à-vis Fungi/Metazoa, the options tried being:
  - 1: Fungi/Metazoa branching first, then Viridiplantae+Kinetoplastida and Alveolata
  - 11: Viridiplantae+Kinetoplastida branching first, then Alveolata and Fungi/Metazoa
  - 12: Alveolata branching first, then Viridiplantae+Kinetoplastida and Fungi/Metazoa

[458] Cetartiodactyla in the dataset used are even-toed ungulates (e.g., *Bos taurus* - cattle).

Subset 8: Some Eukaryota

In the second round, subset 8 was able to distinguish between the most altered trees[459]. It had 8862 amino acids, in 25 proteins (with ADH1 counted as 1). Runs with it were for 200000 generations (2000 samples); two values were used for the burnin used for sump and sumt, as shown below:

Phylogeny Tested	Burnin=1000		Burnin=1800	
	Arith. M.	Harmon. M.	Arith. M.	Harmon. M.
1 (orig):	**-101,987.15**	**-112,295.99**	**-101,987.15**	**-102,403.78**
2:	-113,788.62	-130,388.44	Not done[460]	Not done
9:	-105,504.75	-117,183.08	-105,504.75	-106,242.48
10:	-113,836.44	-122,585.23	Not done	Not done
11:	**-89,957.77**	-121,402.46	**-89,957.77**	**-90,645.00**
12:	-98,529.80	**-108,336.75**	-98,529.80	-98,919.57

Of the above, 11 and 12 are both concerned with alterations in the relative positions of Kinetoplastida (considered at the time to be with Viridiplantae) and Alveolata vis-à-vis Fungi/Metazoa[461]. The inconsistent behavior of their probabilities, as well as the error - in hindsight - of leaving out Viridiplantae from

[459] Note, incidentally, that while this is an indicator of "power" in terms of variety of species, it is not necessarily an indicator of how valid the results are likely to be, since it could be accompanied by a low number of amino acids or other problems (e.g., lack of overlap between amino acids from different species). Indeed, the functionality of REC-I-DCM3 (see "Species subsets", on page 101) argues that a dataset with a wide variety (as opposed to a large number) of species may be *less* likely to give valid results. It is also, of course, not an indication of "power" if some variety of copying error took place in the (manual) construction of the variations; this was unfortunately the case with several of the variants, some of which were not noted as such until after the run.

[460] In some cases, the use of a greater "burnin" value (discarding more of the initial samples) was not done. In most cases, this was if the log probability was very uncertain or dropped at the end of the run, but in some cases because it was obvious from the graph of probabilities given by MrBayes that all of the probability values were worse than those from at least one contradictory phylogeny.

[461] Note that the trees with "non-Fungi/Metazoa Eukaryota only shown" have "Fungi" and "Metazoa" substituted for the groups of species in question - this is for display purposes only, since these groups were *not* made into a composite sequence (see "Further sequence processing: Group sequence creation", on page 96) for these runs.

the subset[462], are why these results were not used. Similarly, 9 and 10 are for differing positions of Cetartiodactyla with respect to Primates and Rodentia; the later conclusion of Rodentia as early branching off the placental clade (see “Tree search with Mammalia (subset)”, on page 316) - an alternative that was unfortunately not tried - distinctly reduces the significance of these results. For 1 versus 2, the species moved were *D. discoideum* and *Candida oleophila*. Please see the trees below, on pages 269-283.

[462] Admittedly, given that it was later concluded that Viridiplantae were basal among Eukaryotes, and at the time of subset 8’s creation they was considered to be together with Kinetoplastida, it is possible that no real information would have been gained by such an expanded subset. Given this, and that the results from a burnin of 1800 were consistently in favor of arrangement 11 (in contradiction to the final tree arrangement from the tree search), further exploration of this area is recommended.

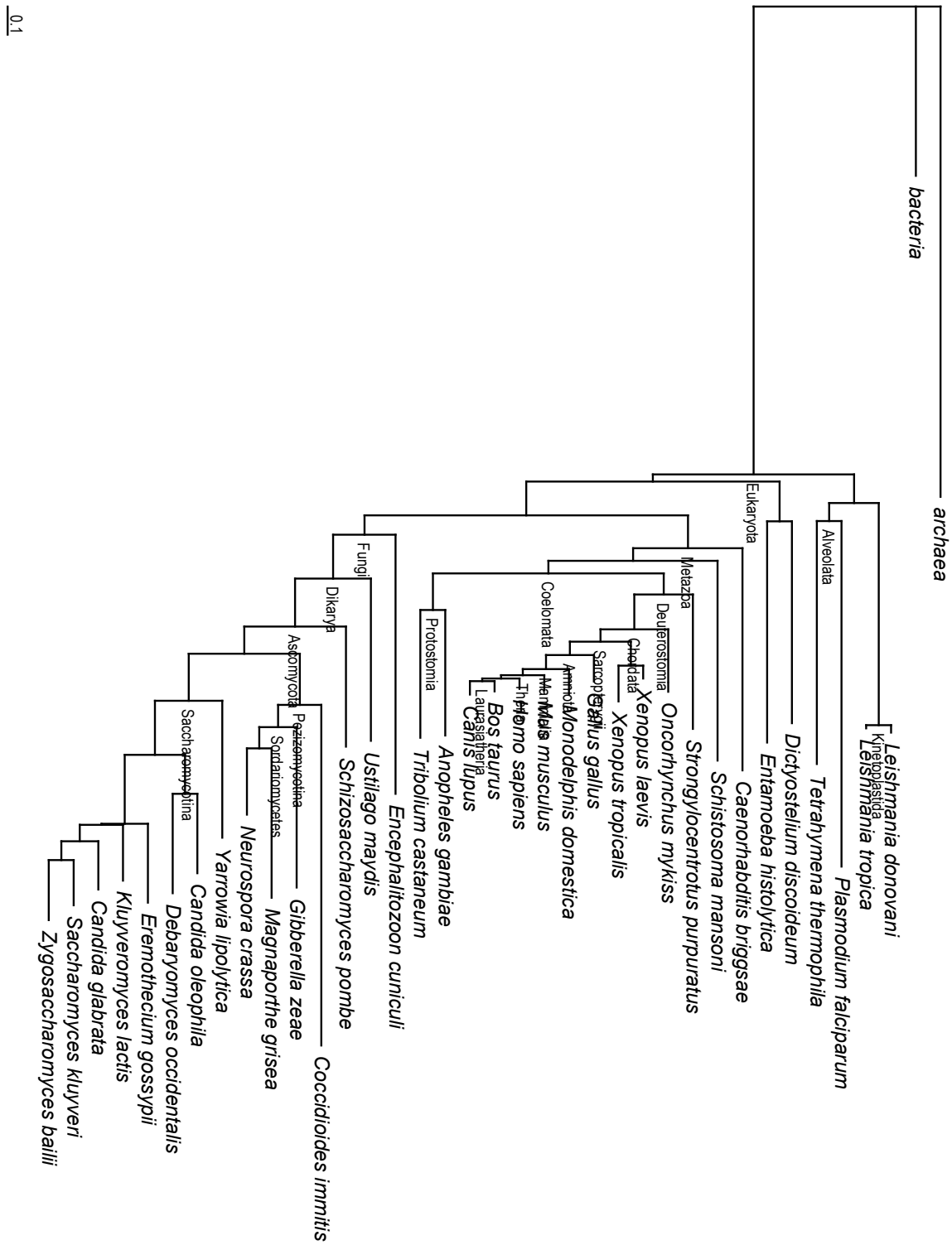


Figure 4.T.r2.s8.c.p: Round 2 subset 8 of final tree, phylogram

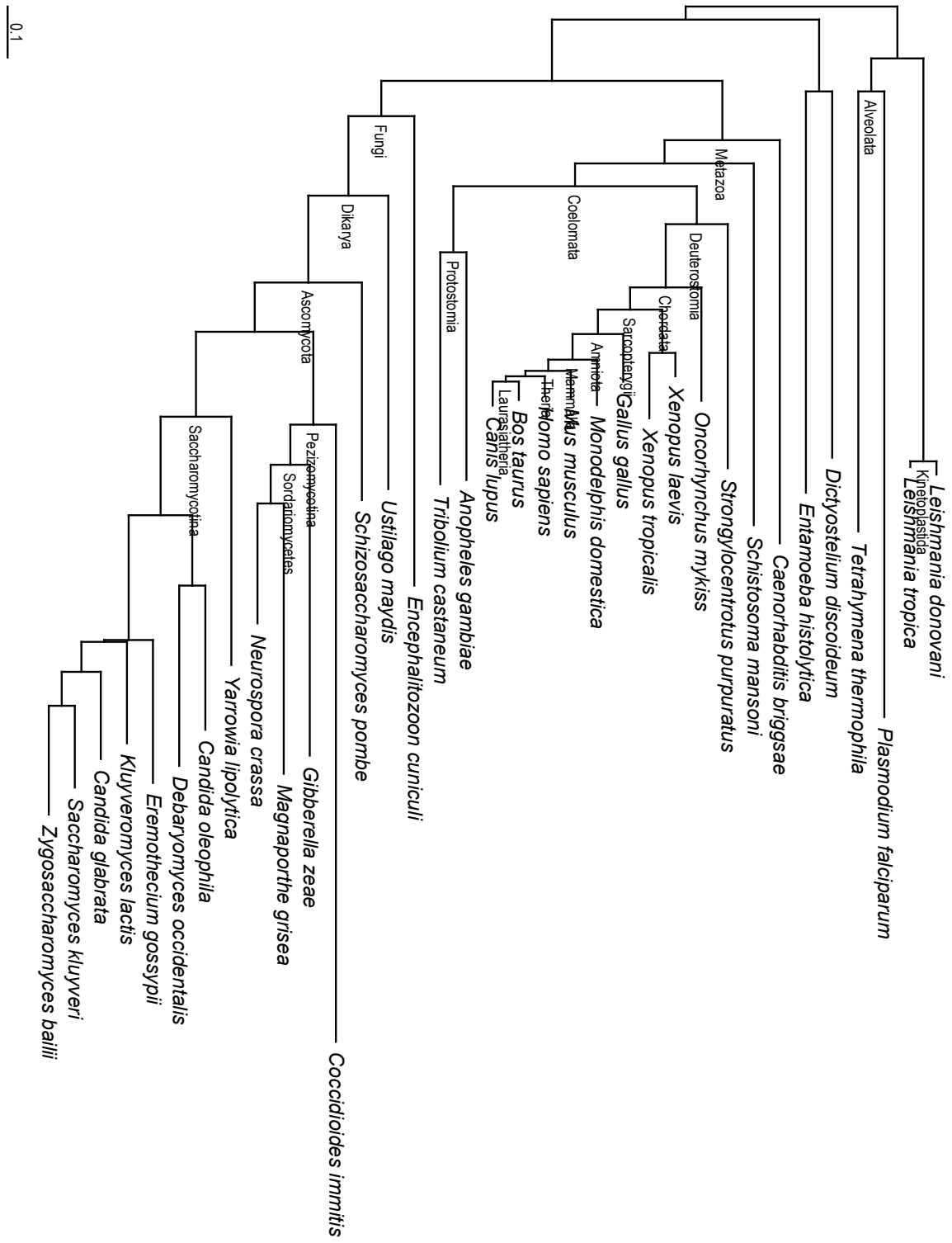


Figure 4.T.r2.s8.c.p.eukaryota: Round 2 subset 8 of final tree, Eukaryota only shown, phylogram

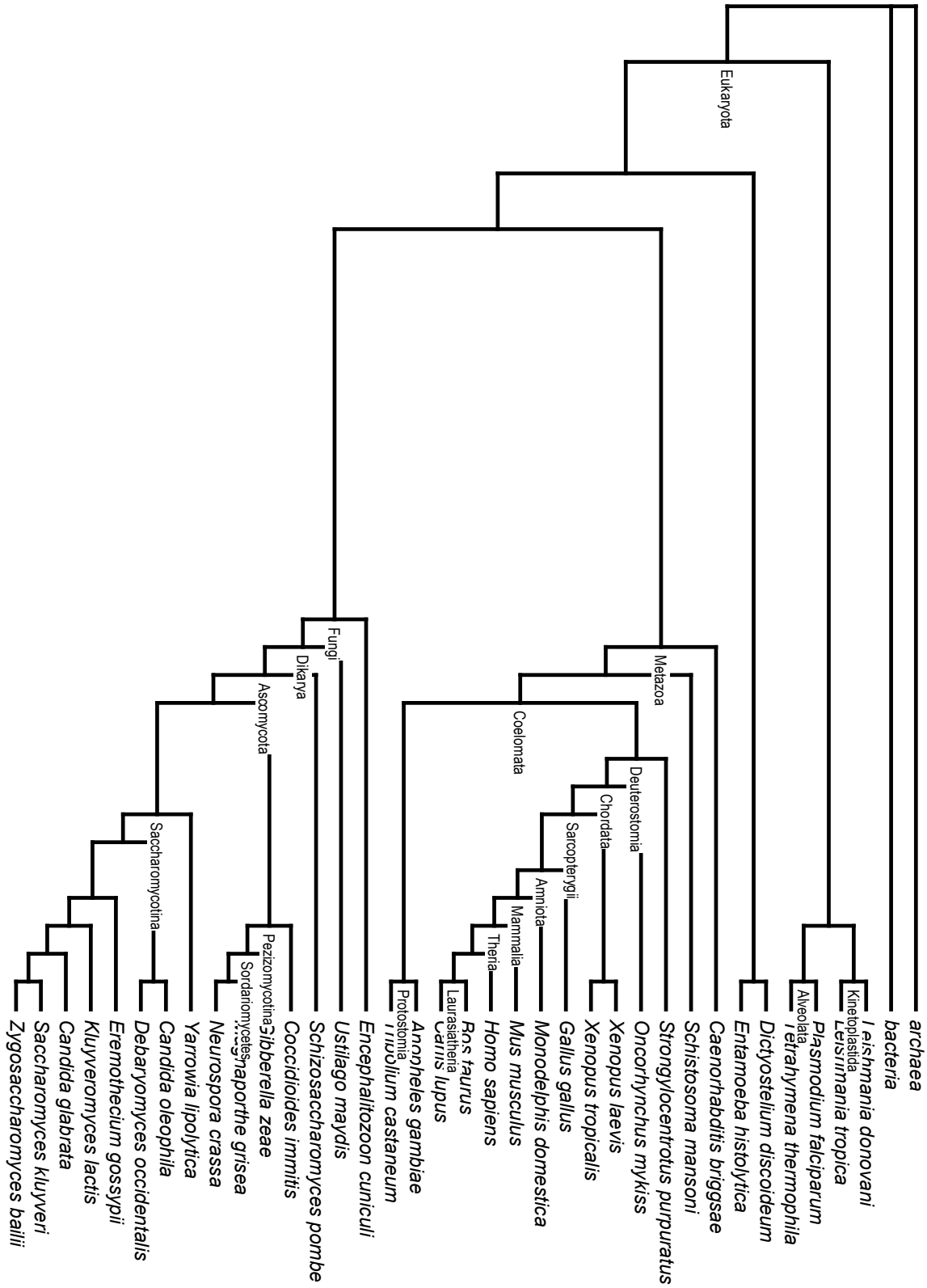


Figure 4.r2.s8.c.c: Round 2 subset 8 of final tree, cladogram

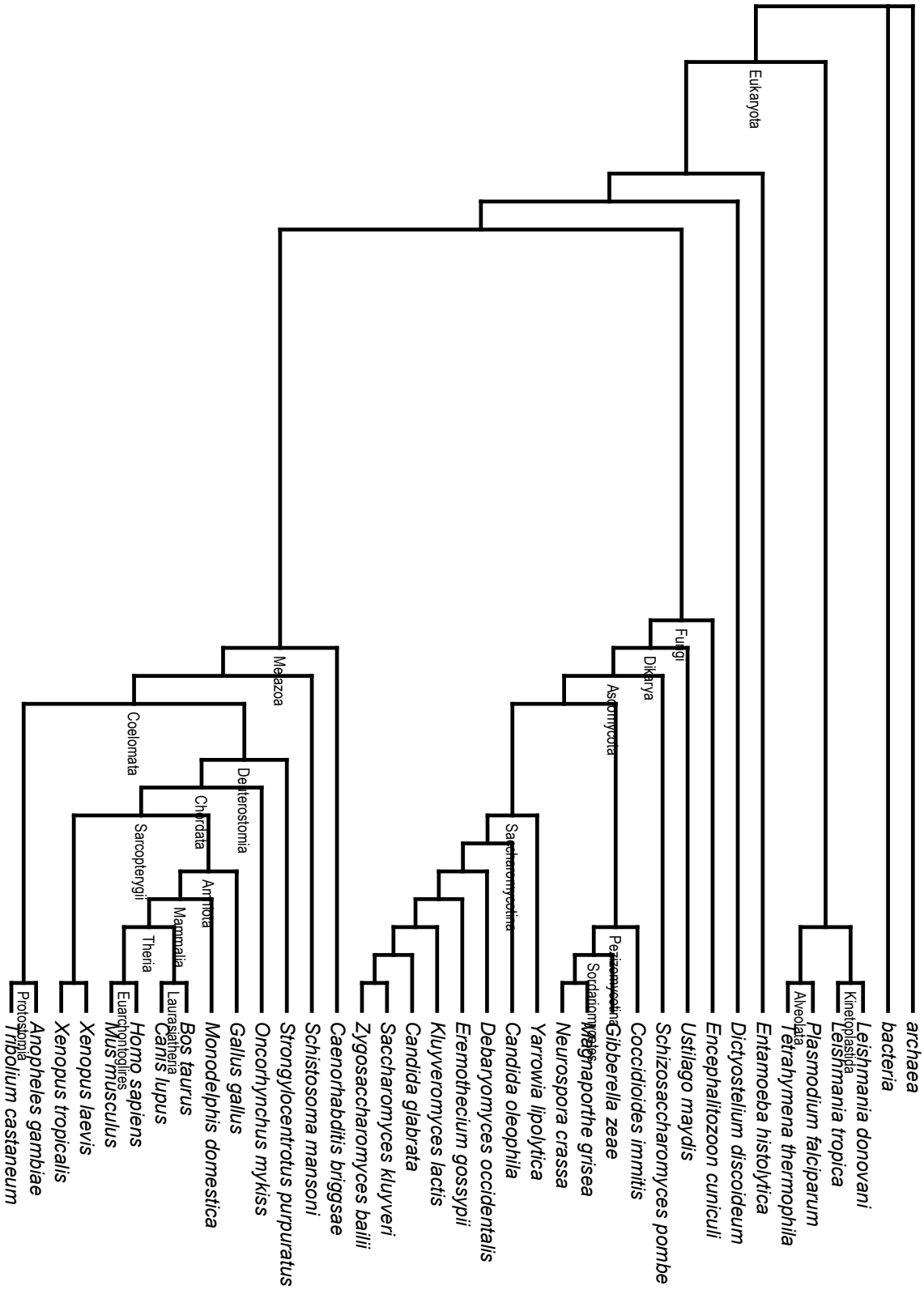


Figure 4.T.r2.s8.1: Round 2 subset 8, original (tree 1) arrangement, cladogram

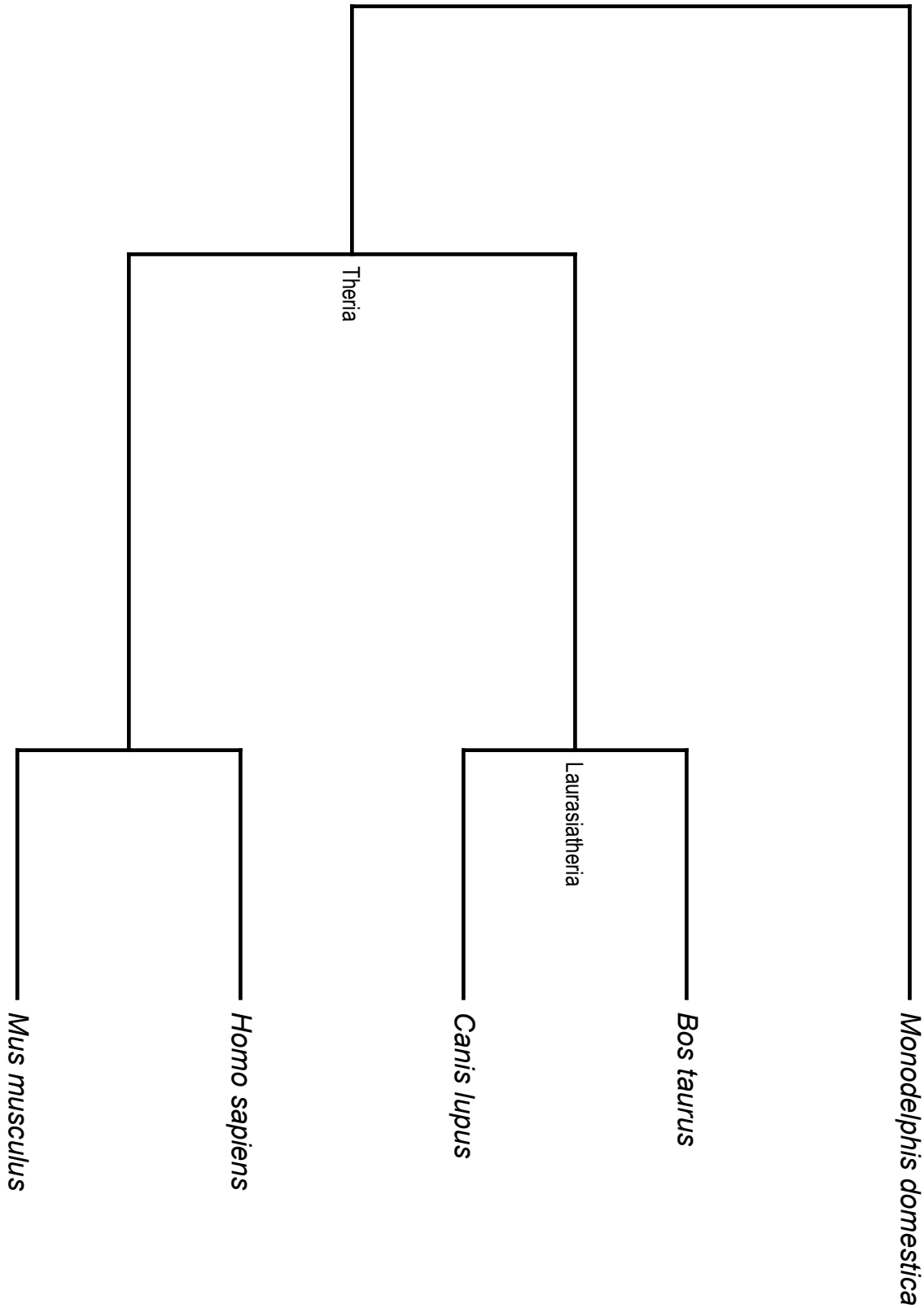


Figure 4.T.r2.s8.1.mammalia: Round 2 subset 8, original (tree 1) arrangement, Mammalia only shown, cladogram

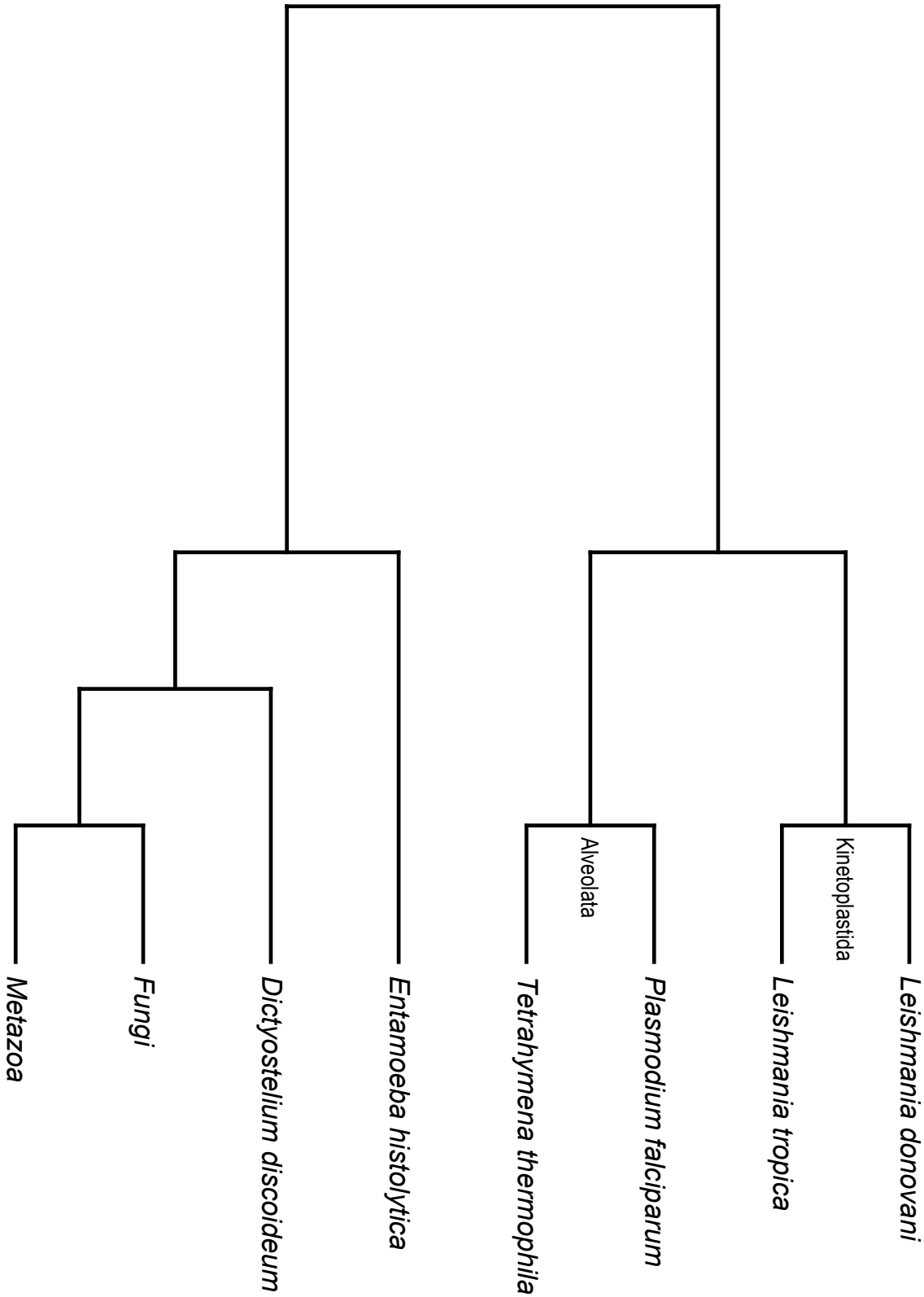


Figure 4.T.r2.s8.1.nfm: Round 2 subset 8, original (tree 1) arrangement, non-Fungi/Metazoa Eukaryota only shown, cladogram

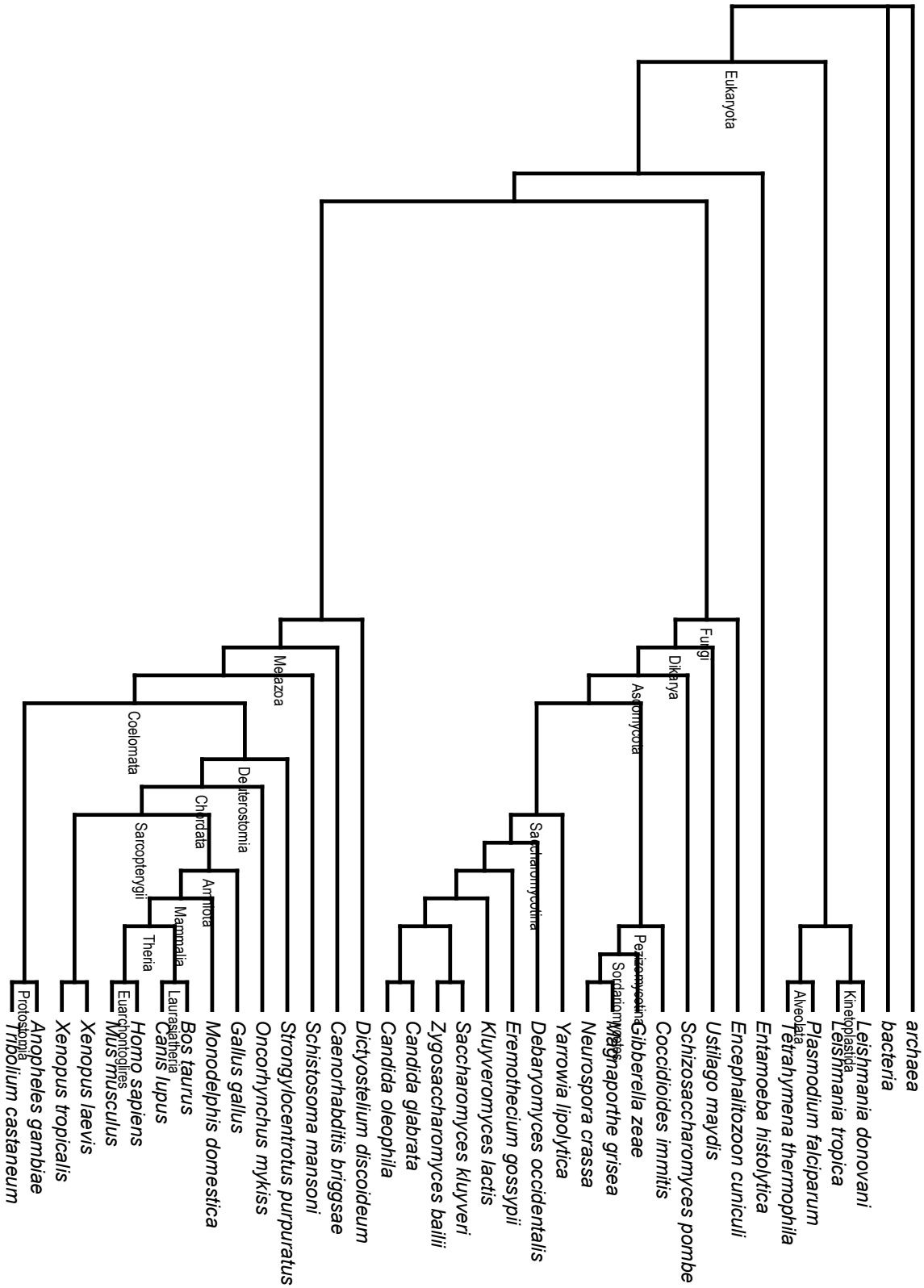


Figure 4.T.r2.s8.2: Round 2 subset 8, tree 2 arrangement, cladogram

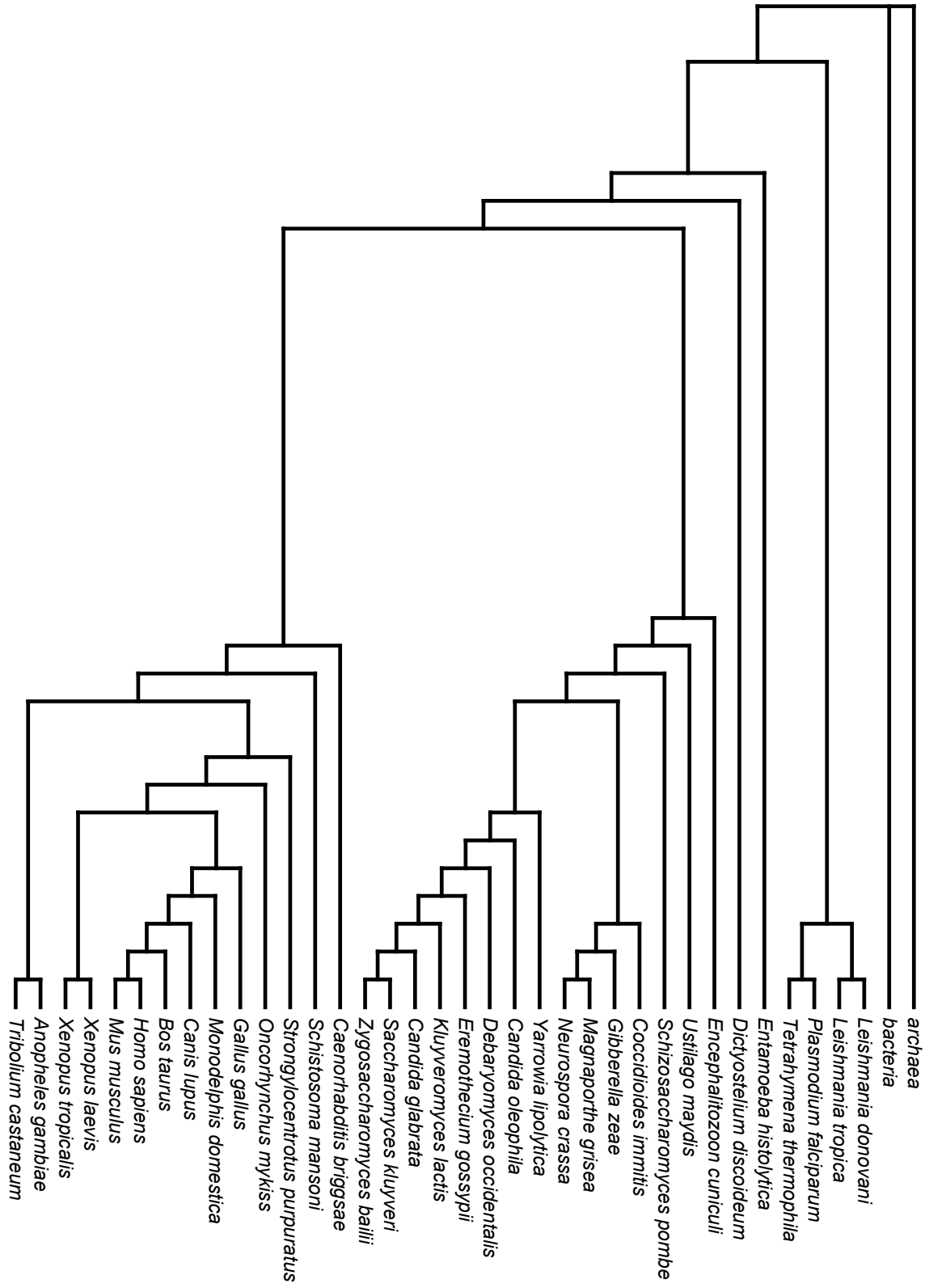


Figure 4.T.r2.s8.9: Round 2 subset 8, tree 9 arrangement, cladogram

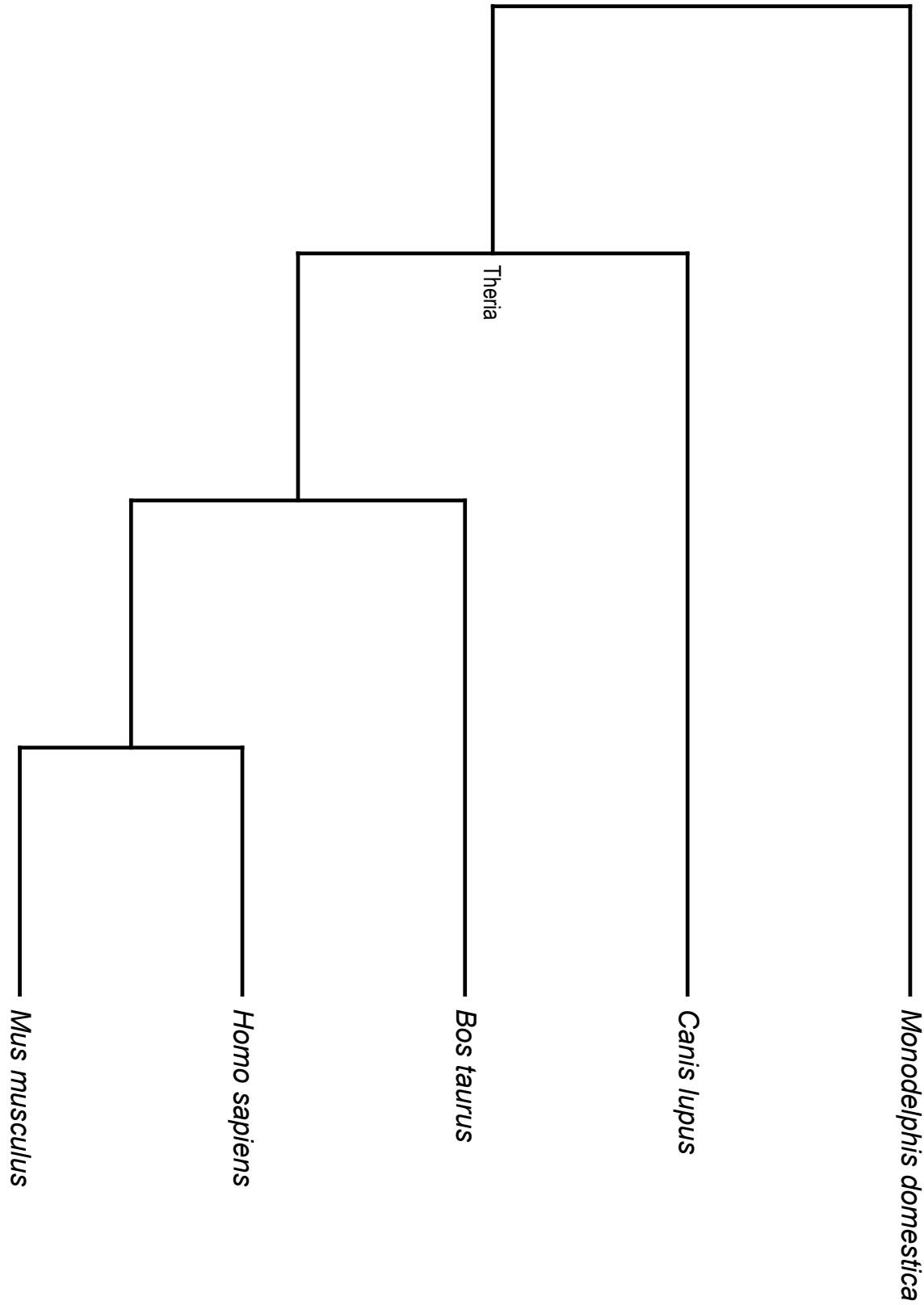


Figure 4.T.r2.s8.9.mammalia: Round 2 subset 8, tree 9 arrangement, Mammalia only shown, cladogram

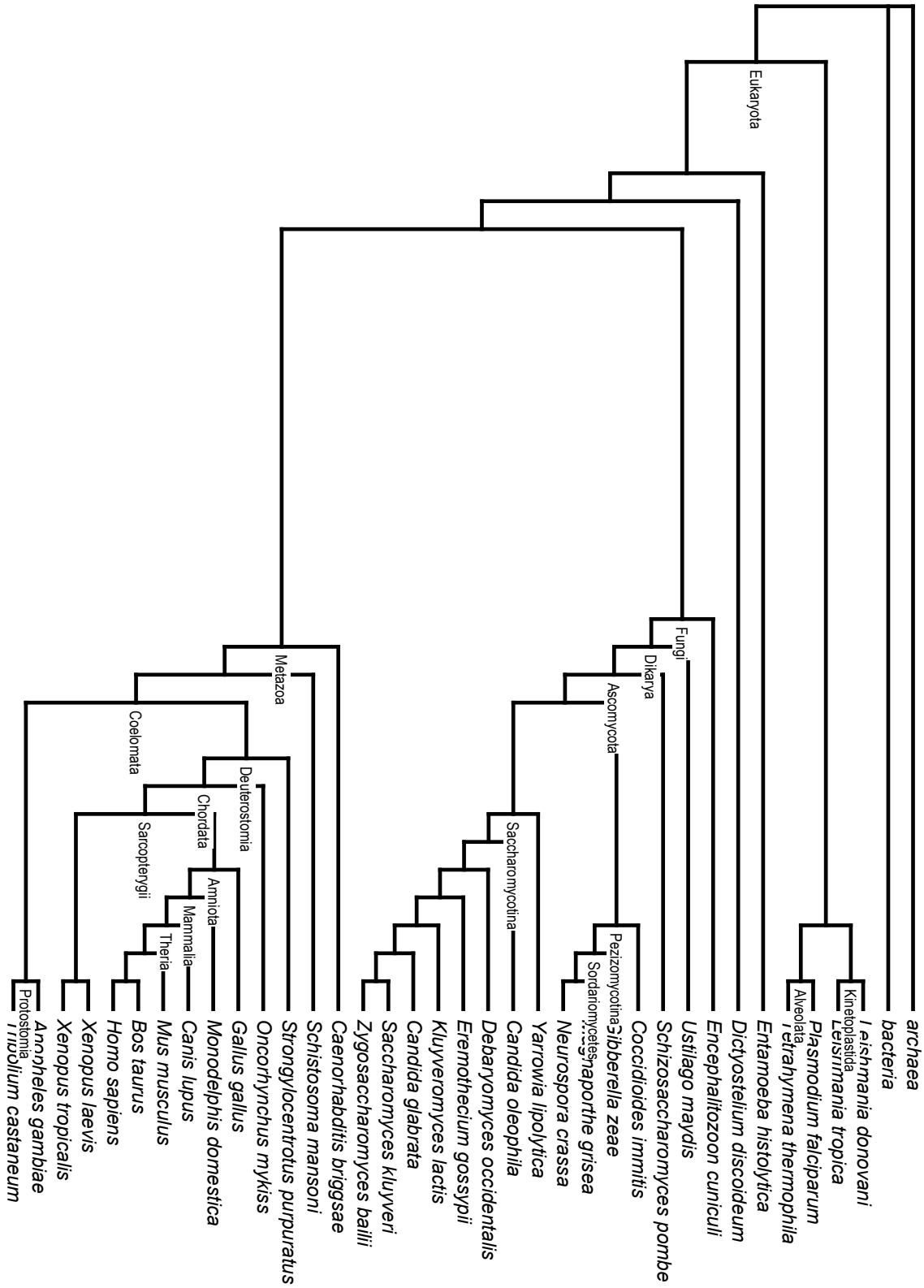


Figure 4.T.r2.s8.10: Round 2 subset 8, tree 10 arrangement, cladogram

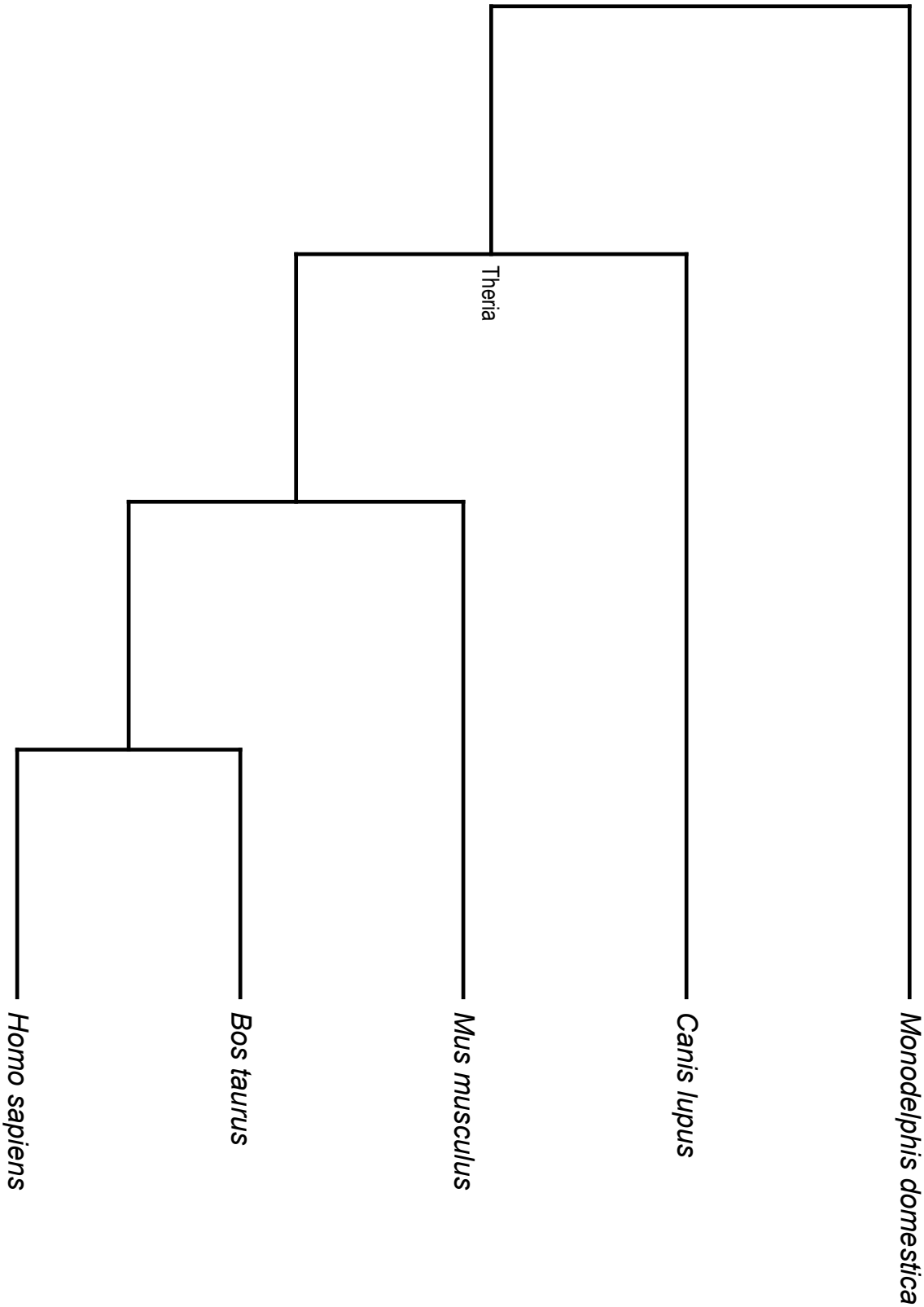


Figure 4.T.r2.s8.10.mammalia: Round 2 subset 8, tree 10 arrangement, Mammalia only shown, cladogram

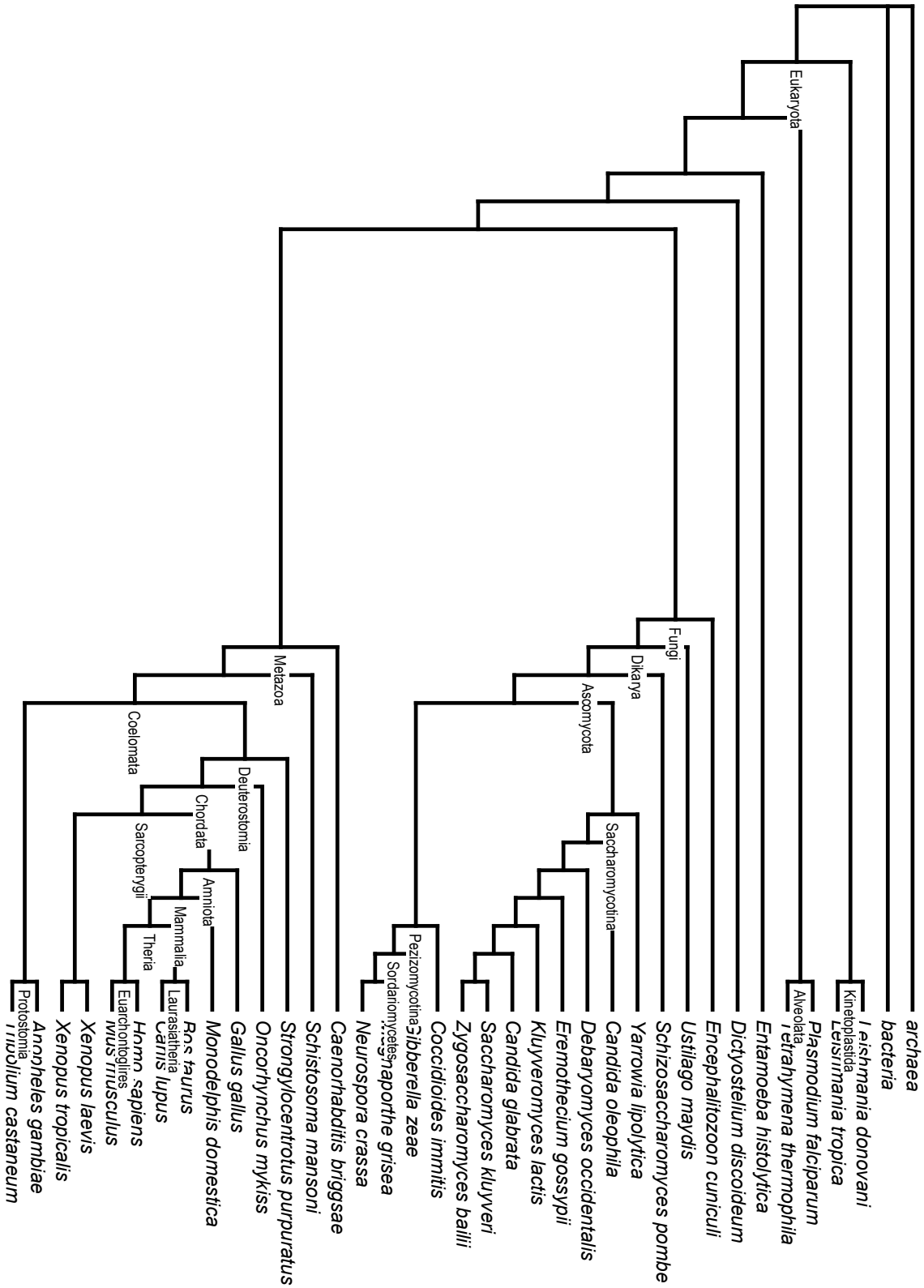


Figure 4.T.r2.s8.11: Round 2 subset 8, tree 11 arrangement, cladogram

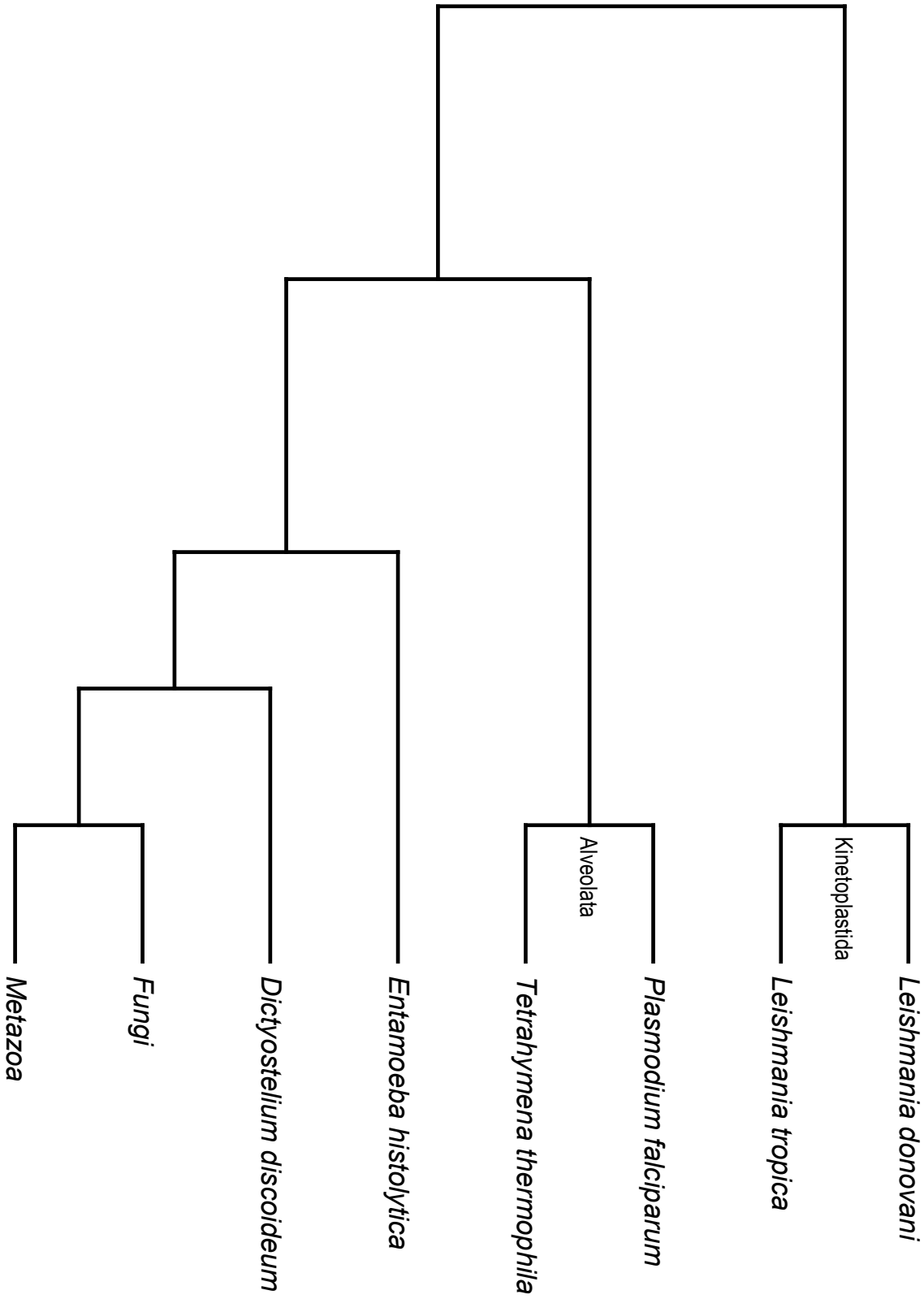


Figure 4.T.r2.s8.11.nfm: Round 2 subset 8, tree 11 arrangement, non-Fungi/Metazoa Eukaryota only shown, cladogram

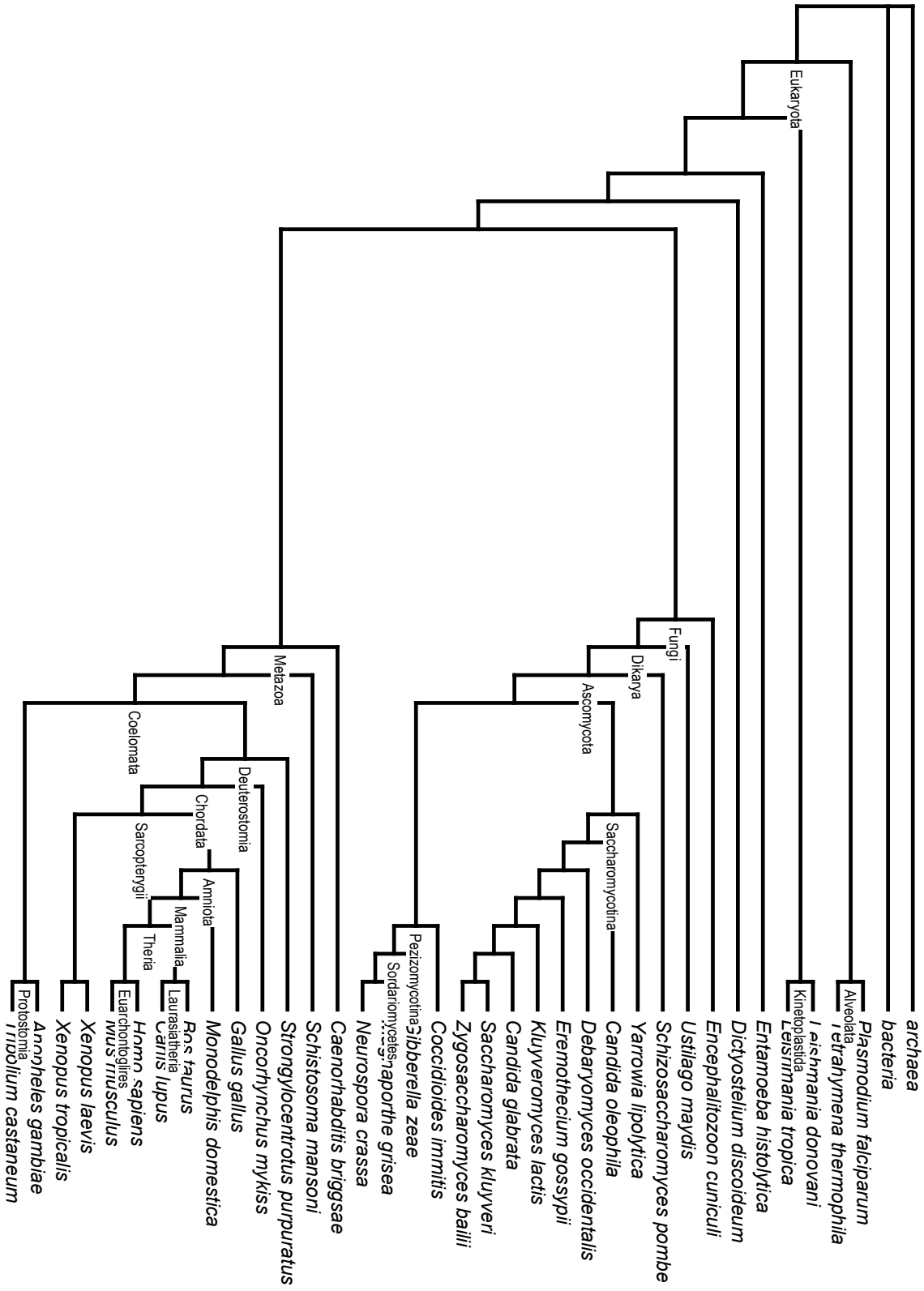


Figure 4.T.r2.s8.12: Round 2 subset 8, tree 12 arrangement, cladogram

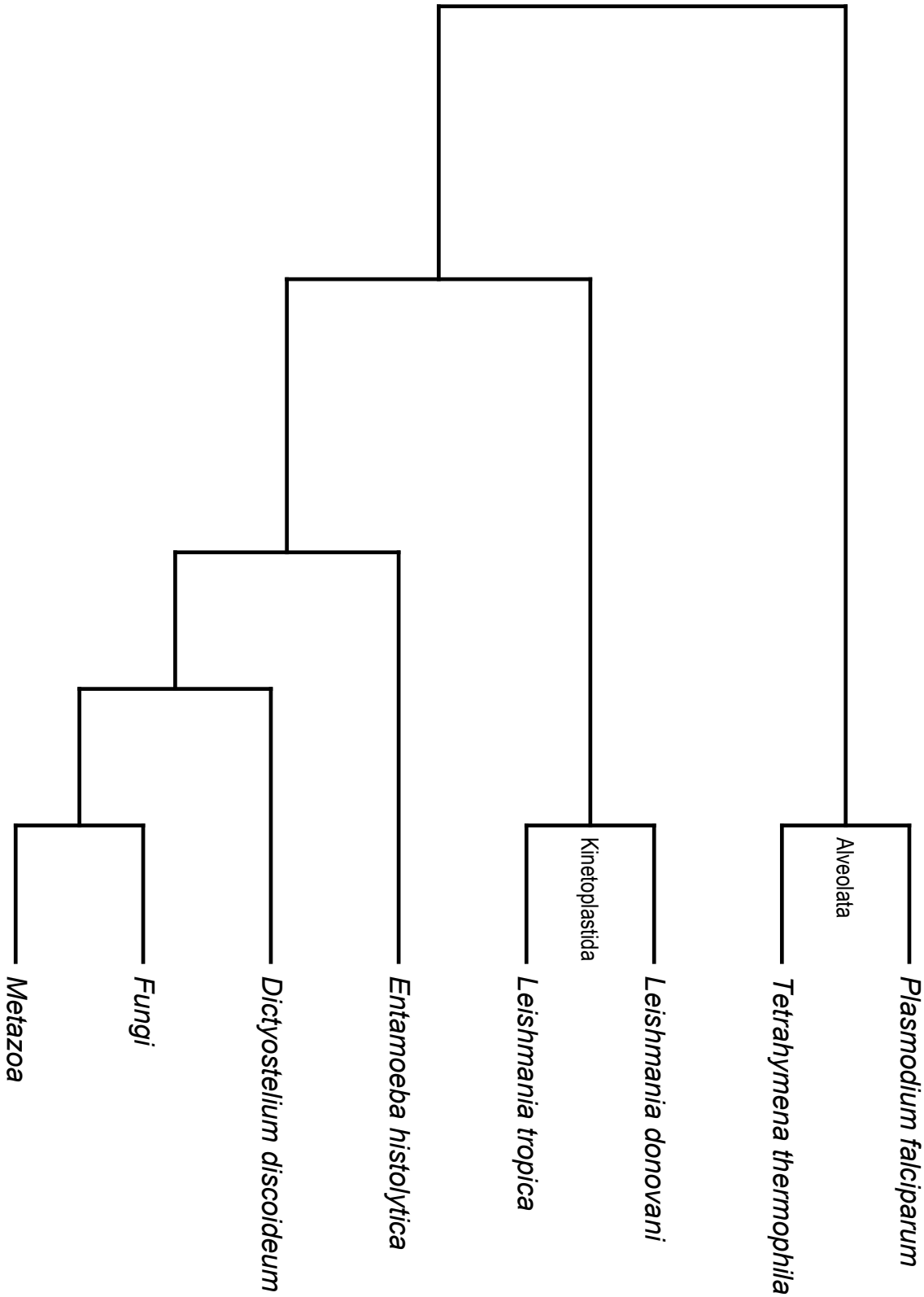


Figure 4.T.r2.s8.12.nfm: Round 2 subset 8, tree 12 arrangement, non-Fungi/Metazoa Eukaryota only shown, cladogram

Subset 10: Some Eukaryota

Subset 10, with 7178 amino acids and 19 proteins (considering ADH1 as 1 protein), was used for runs with 200000 generations (2000 samples), with burnins as given in the table below:

Phylogeny Tested	Burnin=1000		Burnin=1933	
	Arith. M.	Harmon. M.	Arith. M.	Harmon. M.
1 (orig):	**-78,513.21**	**-80,010.81**	**-78,512.42**	**-78,605.04**
2:	-89,148.81	-89,767.73	Not done	Not done
3:	-79,625.34	-80,243.22	-79,624.78	-79,625.41
11:	**-75,296.99**	**-75,867.30**	**-75,294.94**	**-75,295.96**
12:	-80,270.86	-82,522.75	-80,270.37	-80,357.06

In the above, 1 versus 2 versus 3 differ in the position of *D. discoideum* (in 3, it is back to branching together with *E. histolytica*, as per the results later indicated in "Tree search with Non-Fungi/Metazoa Eukaryota", on page 313, and prior research (Bapteste *et al.* 2002)). The relative position of these two species has continued to be uncertain[463]. Please see the trees[464] below, on pages 285-291.

[463] Given that it is also known for *Hartmannella cantabrigiensis* (see footnote 234, on page 113), actin, a well-studied protein in *Dictyostelium discoideum*, is likely to be of interest in firming the position of this species. (In other words, it is likely that, when looking at species with long branch lengths such as *Hartmannella cantabrigiensis* and *Dictyostelium discoideum*, even proteins not found to be divergent to below 65% identity may contain enough data to be useful.)

[464] Note that the tree with "non-Fungi/Metazoa Eukaryota only shown" has "Fungi" and "Metazoa" substituted for the groups of species in question - this is for display purposes only, since these groups were *not* made into a composite sequence (see "Further sequence processing: Group sequence creation", on page 96) for these runs.

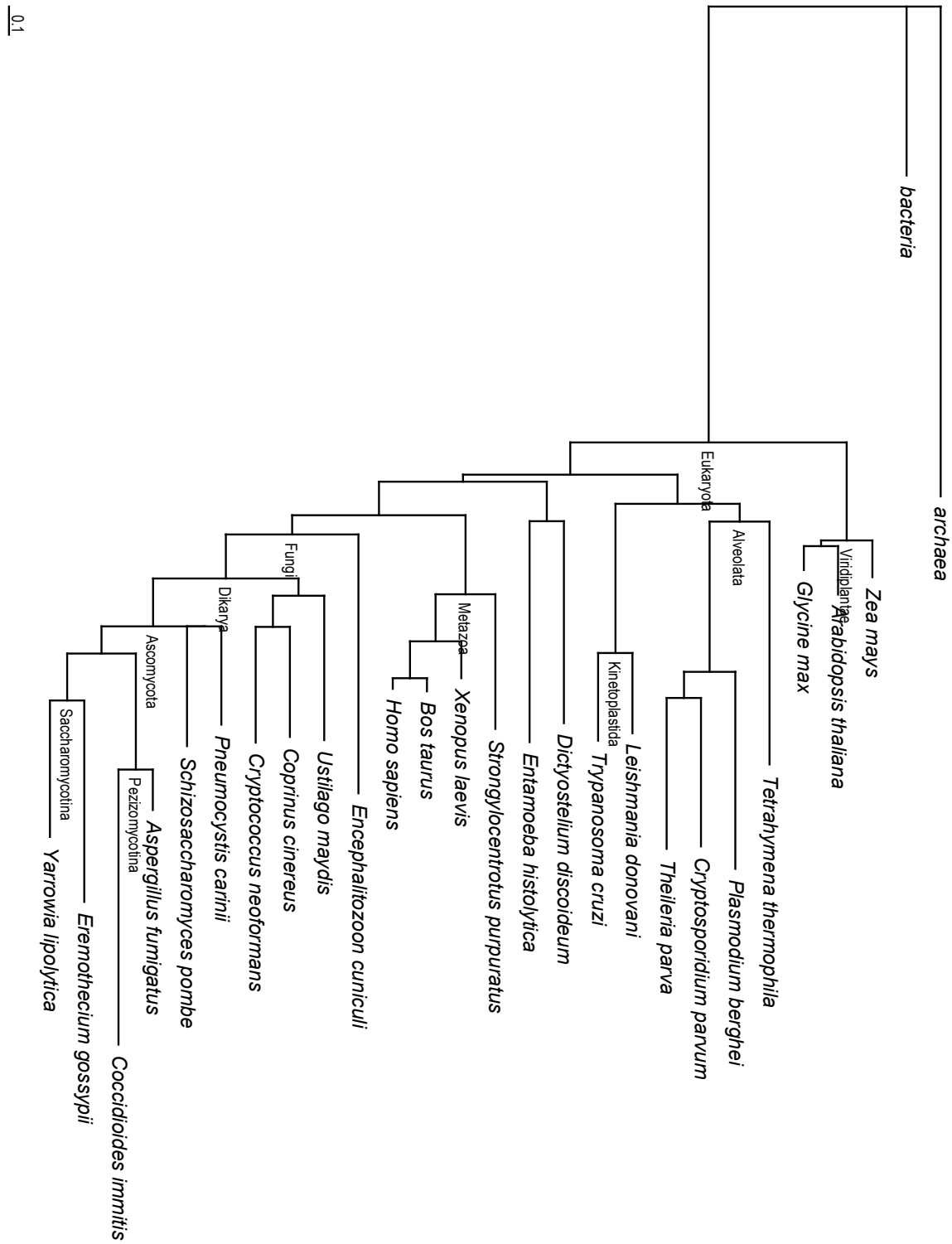


Figure 4.T.r2.s10.c.p: Round 2 subset 10 of final tree, phylogram

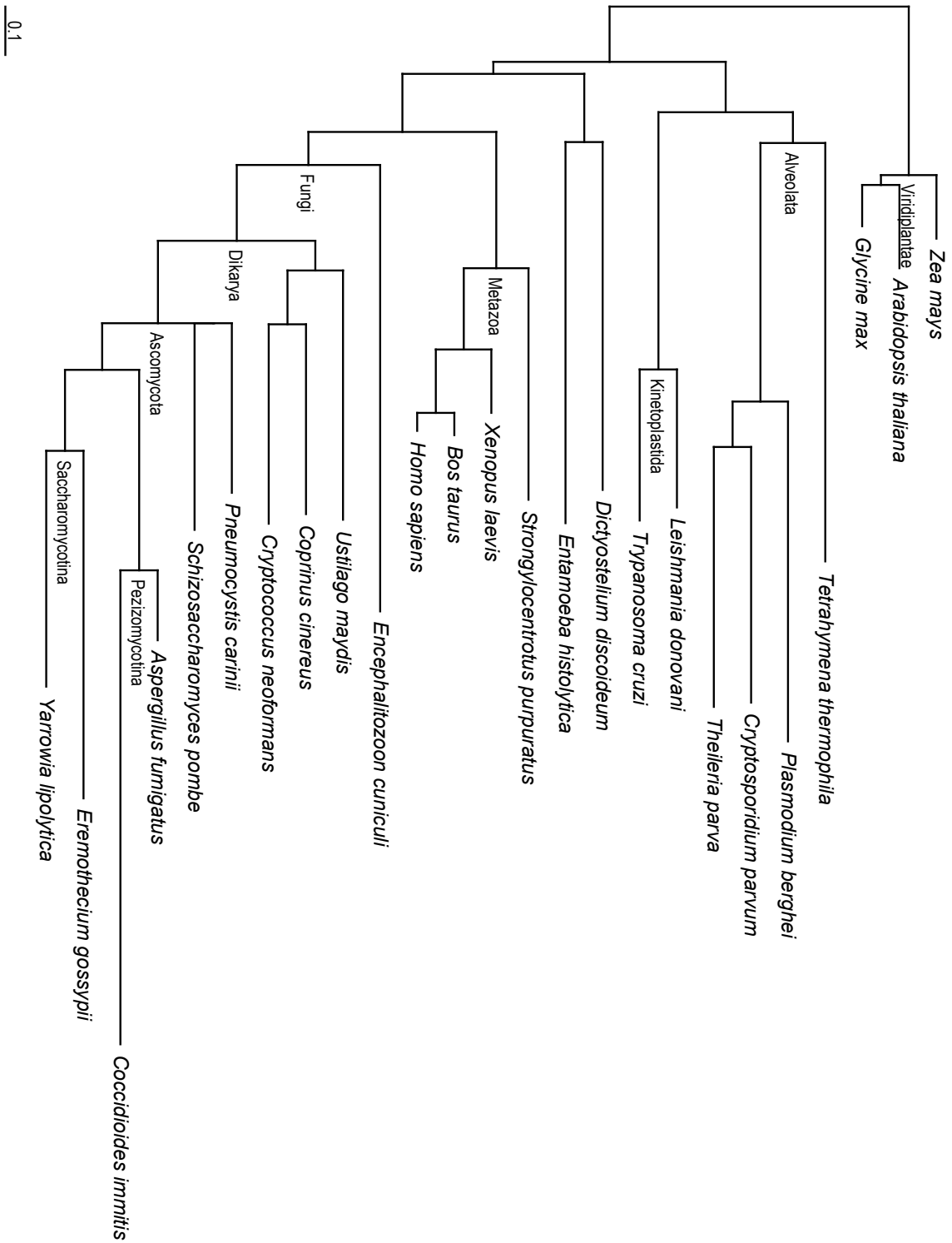


Figure 4.T.r2.s10.c.p.eukaryota: Round 2 subset 10 of final tree, Eukaryota only shown, phylogram

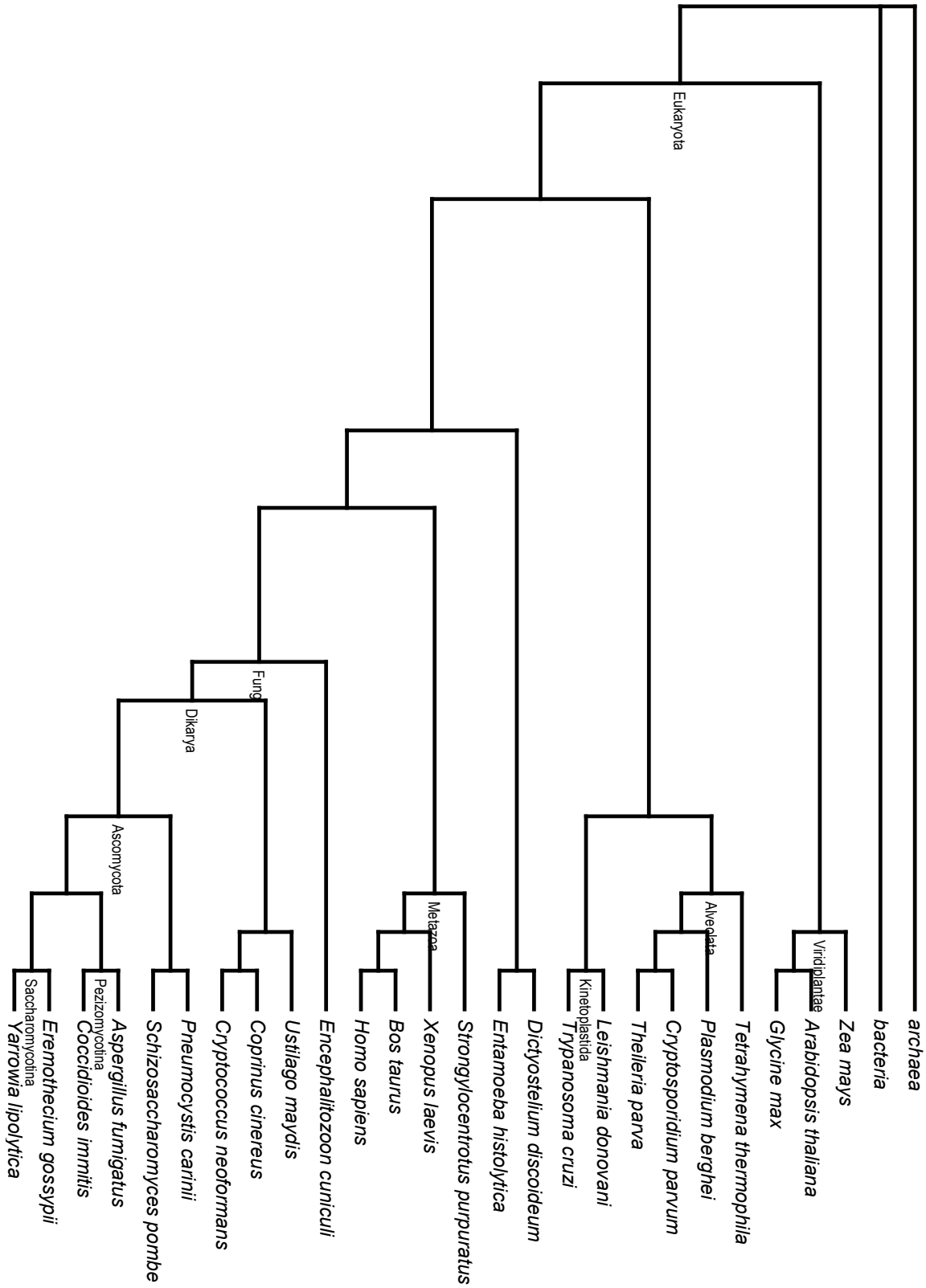


Figure 4.T.r2.s10.c.c: Round 2 subset 10 of final tree, cladogram

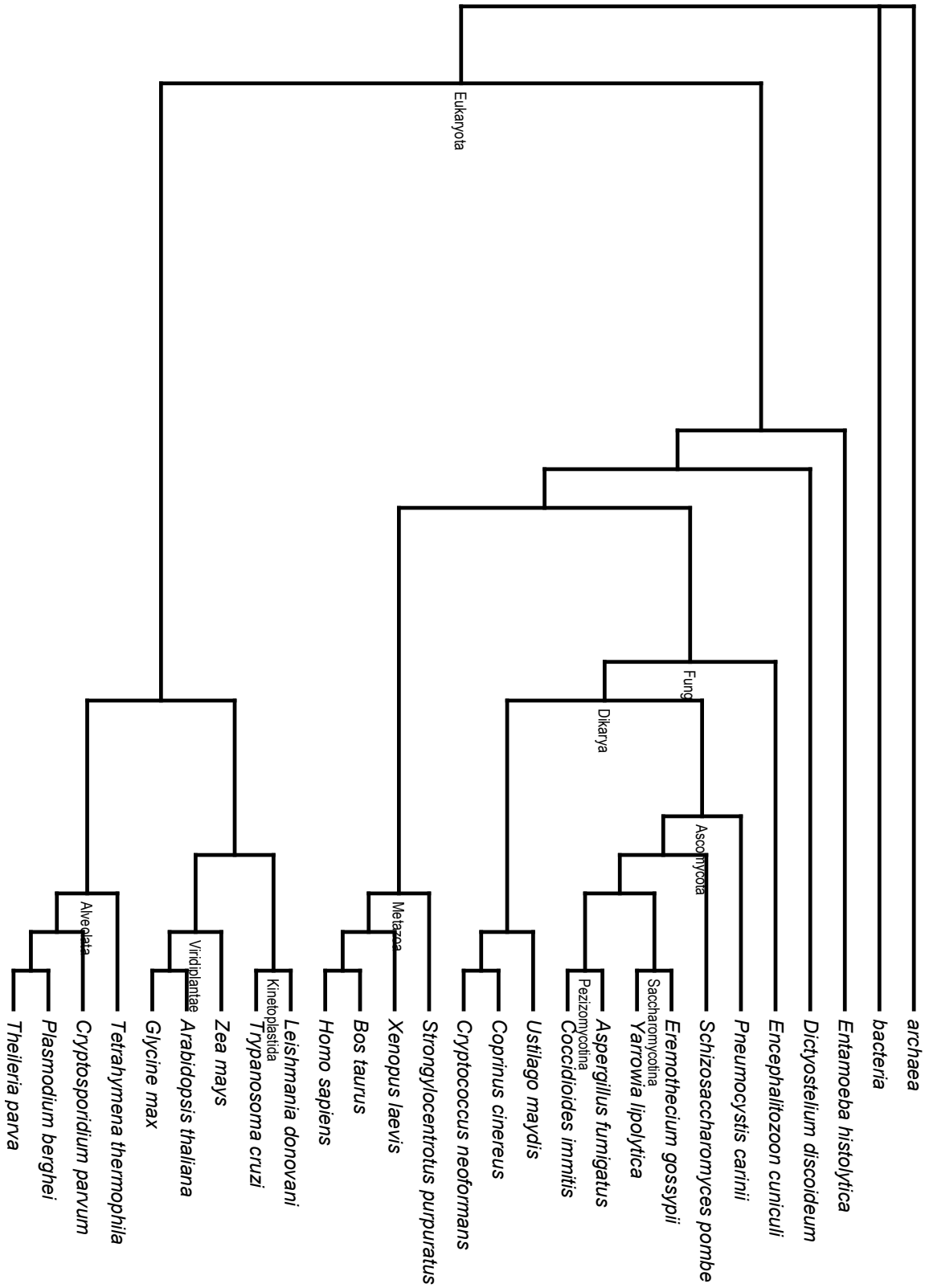


Figure 4.T.r2.s10.1: Round 2 subset 10, tree 1 (original) arrangement, cladogram

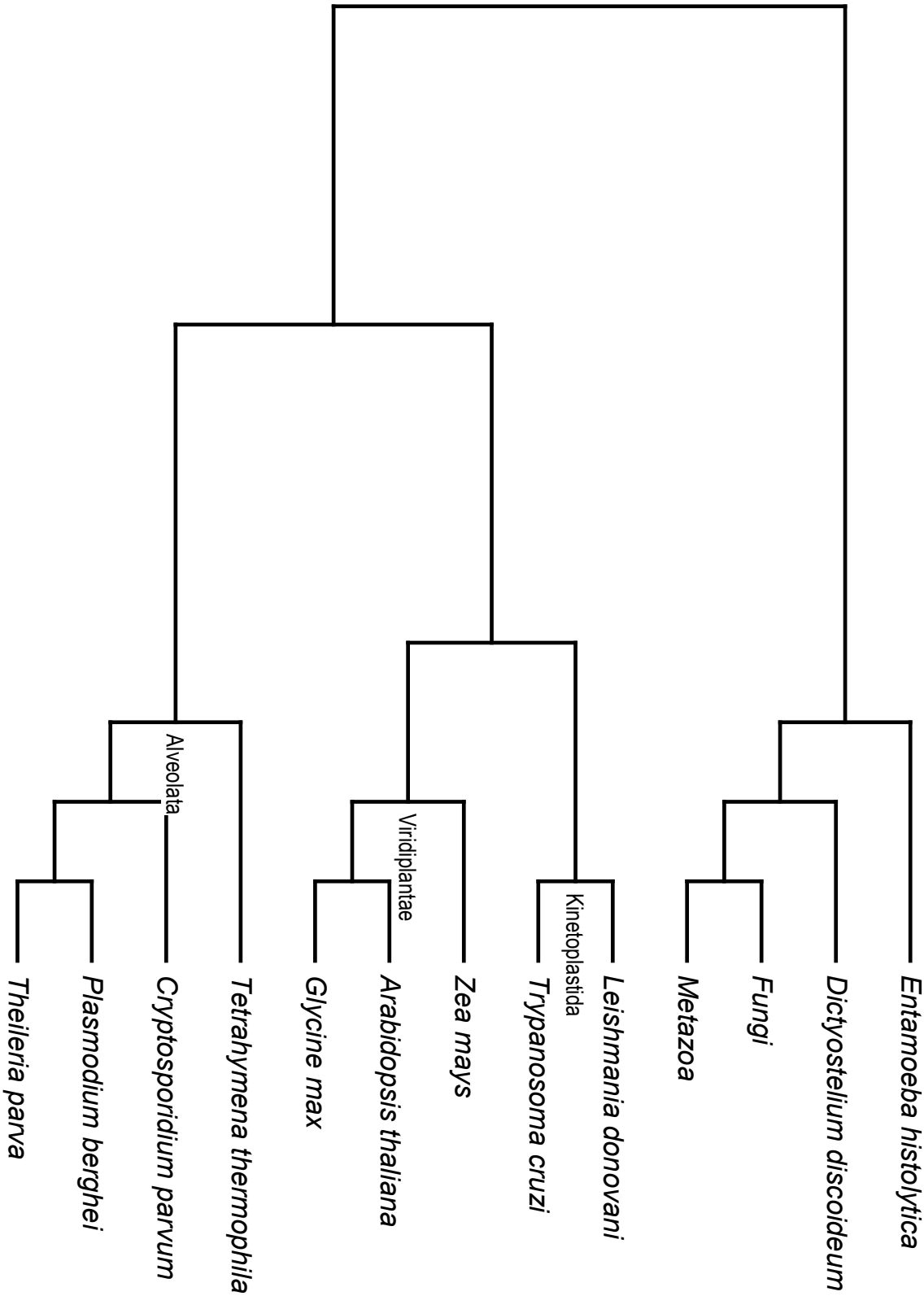


Figure 4.T.r2.s10.1.nfm: Round 2 subset 10, tree 1 (original) arrangement, non-Fungi/Metazoa Eukaryota only shown, cladogram

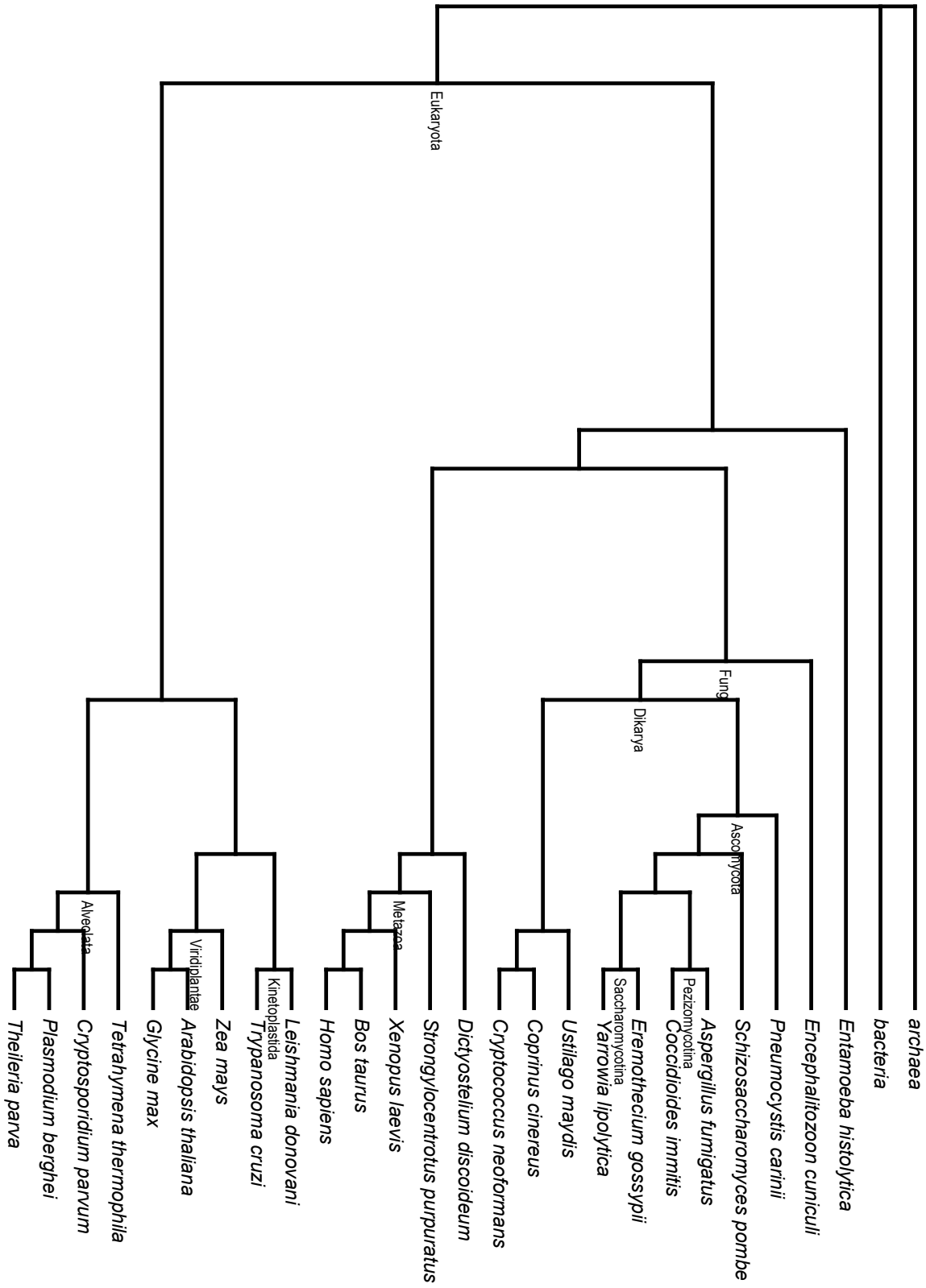


Figure 4.T.r2.s10.2: Round 2 subset 10, tree 2 arrangement, cladogram

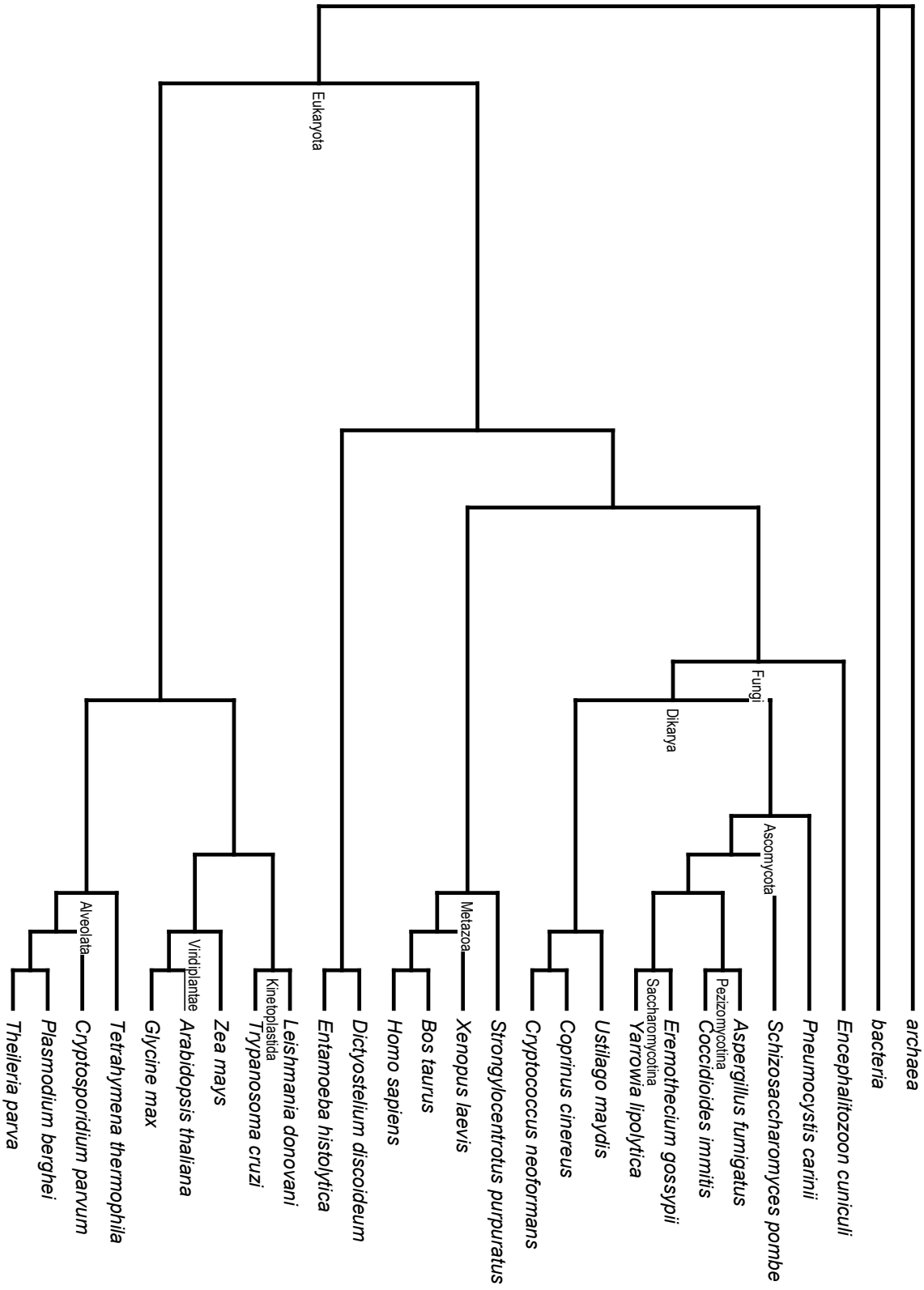


Figure 4.T.r2.s10.3: Round 2 subset 10, tree 3 arrangement, cladogram

Subset 12: Some Eukaryota (Plant/Algae as composite sequence)

Runs of 200000 generations (2000 samples) were also performed with subset 12; this had 8547 amino acids from 24 proteins (considering ADH1 as 1). The burnins were as given in the table below:

Phylogeny Tested	Burnin=1000		Burnin=1916	
	Arith. M.	Harmon. M.	Arith. M.	Harmon. M.
1 (orig):	-117,825.28	**-120,981.67**	-117,824.42	-117,861.71
5:	**-116,944.57**	-127,879.13	**-116,944.54**	**-116,965.70**
11:	-113,870.47	-116,732.52	-113,870.25	-114,057.69
12:	**-106,058.92**	**-109,514.26**	**-106,058.92**	**-106,395.41**

The results from the above for 11 and 12 are, as with subsets 8 and 10, contradictory (to those for other subsets); this was one reason[465] for running a tree search ("Tree search with Non-Fungi/Metazoa Eukaryota", on page 313) focusing on Alveolata, Kinetoplastida, and Viridiplantae. Example trees for this subset are shown on pages 293-297.

[465] Another was the long-branch attraction and other distortions seen in "Tree search with Eukaryota (subset)", on page 300.

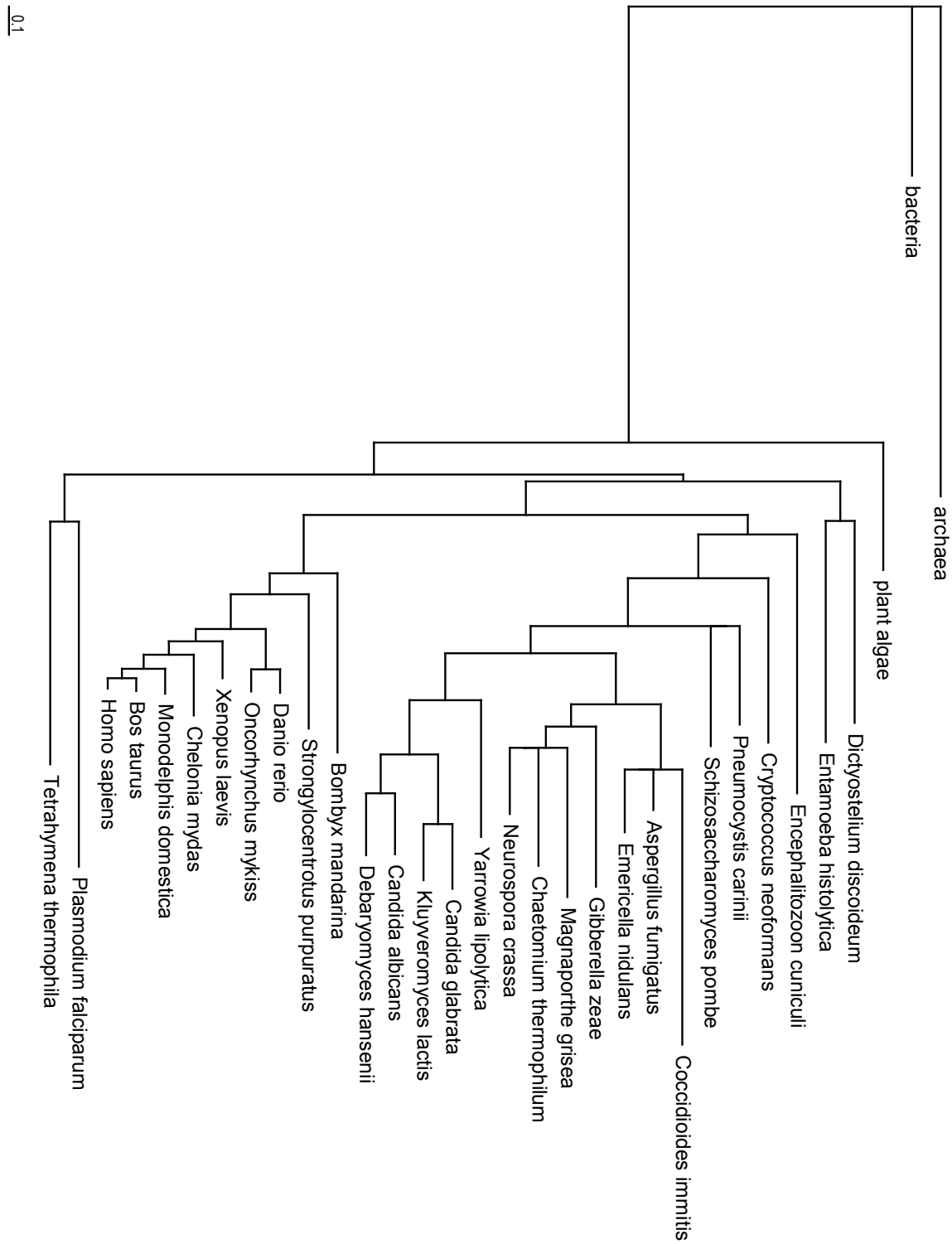


Figure 4.T.r2.s12.c.p: Round 2 subset 12 of final tree, phylogram

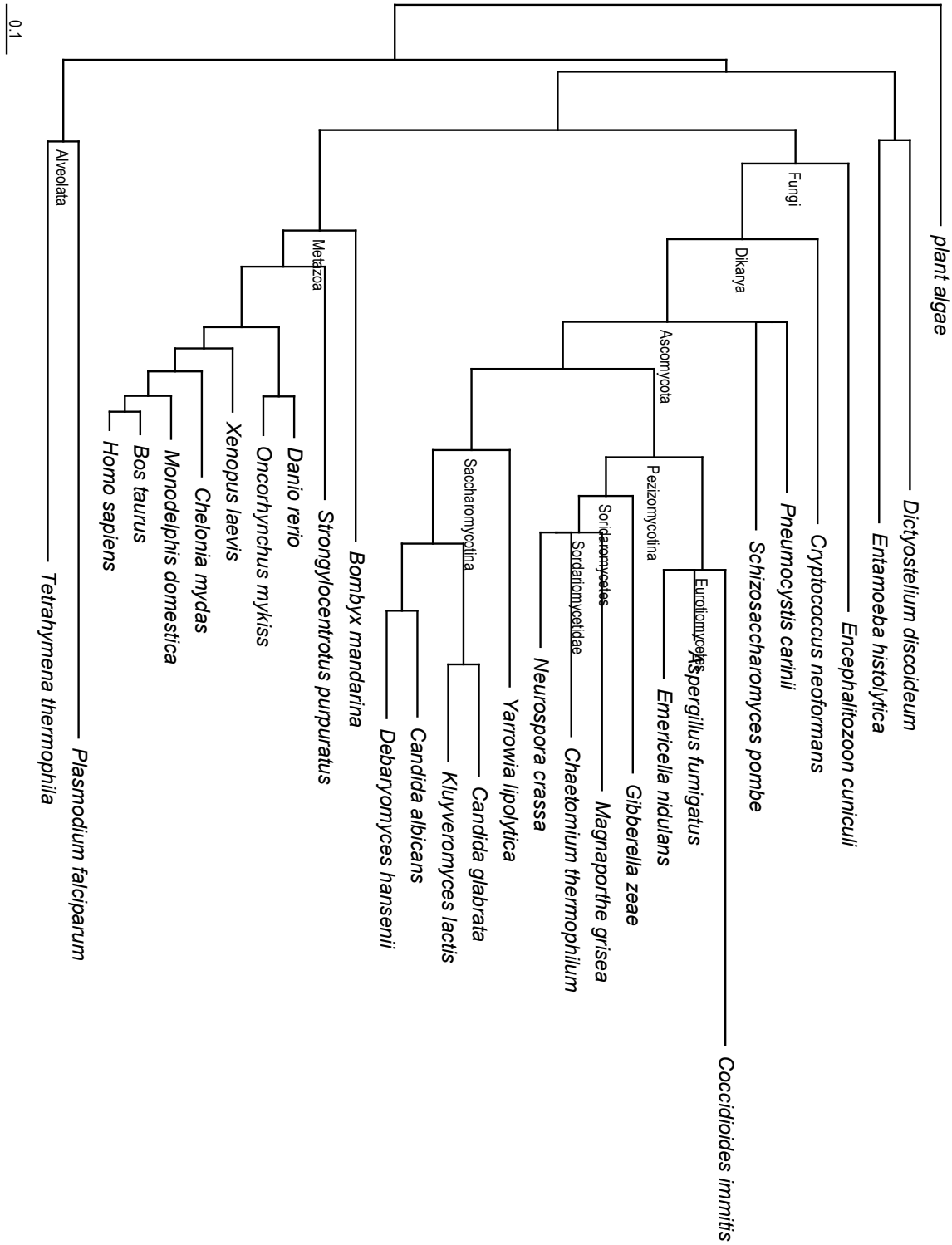


Figure 4.T.r2.s12.c.p.eukaryota: Round 2 subset 12 of final tree, Eukaryota only shown, phylogram

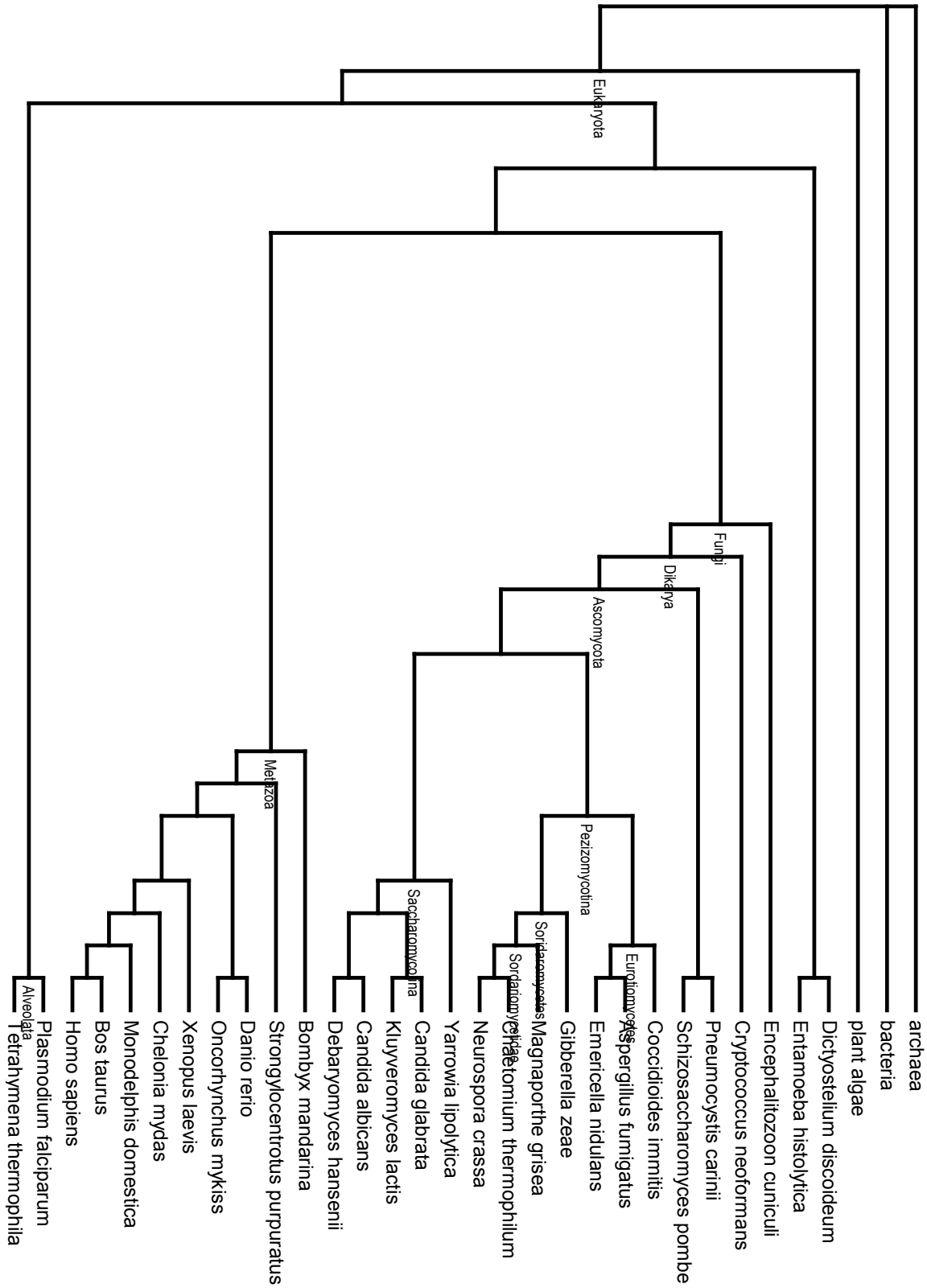


Figure 4.T.r2.s12.c.c: Round 2 subset 12 of final tree, cladogram

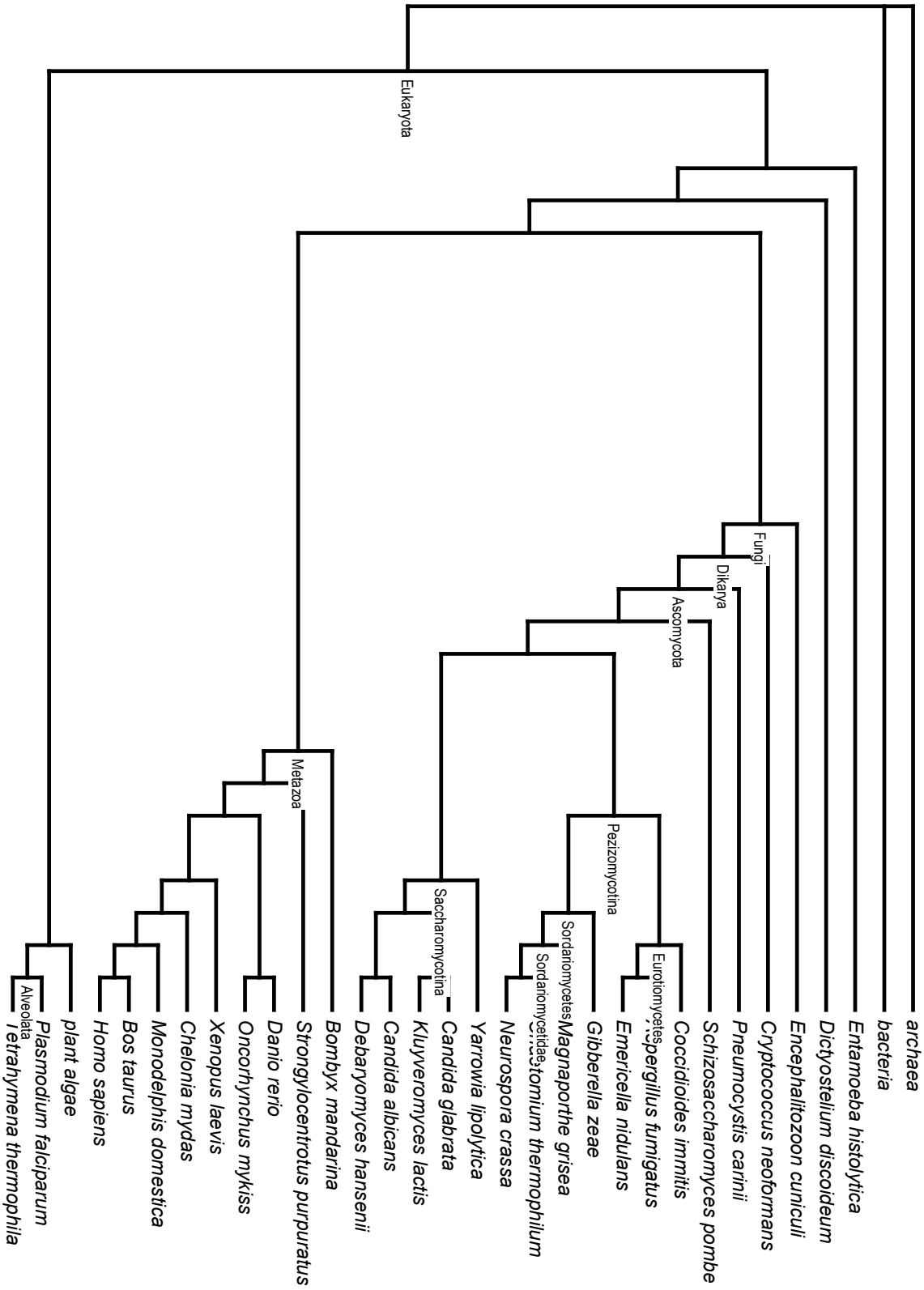


Figure 4.T.r2.s12.1: Round 2 subset 12, original (tree 1) arrangement, cladogram

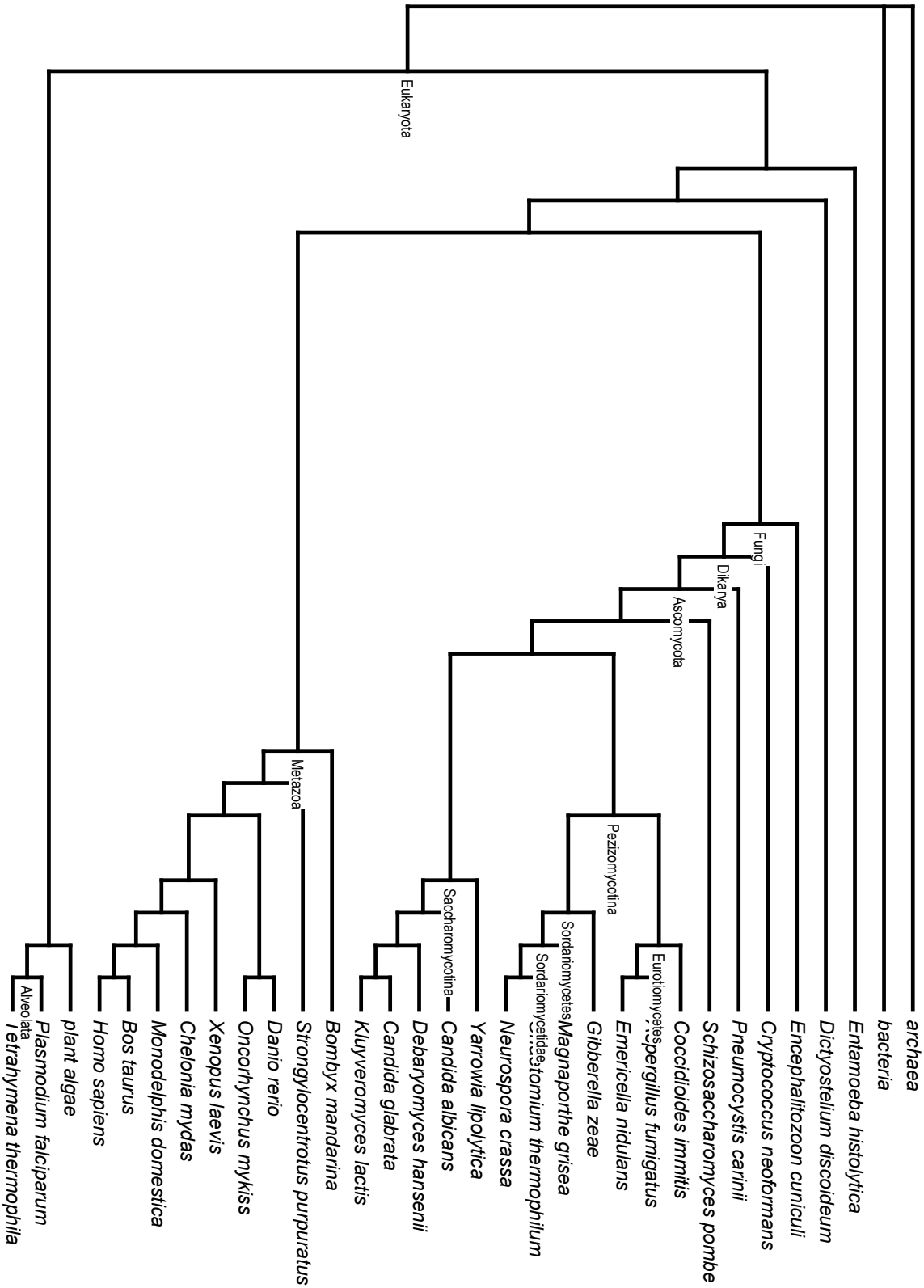


Figure 4.T.r2.s12.5: Round 2 subset 12, tree 5 arrangement, cladogram

Summary of second round results

The below table is a summary of the tree rearrangement (hypothesis) results (see "rearrangements", on) from each subset in this round, with boldface indicating the stronger of two results when applicable:

Subset	1 vs. 2, 5	1 vs. 2 vs. 3	1 vs. 11 vs. 12	1 vs. 9 vs. 10
8	(Not[466] 2)	(Not 2)	**11** or 12	**1**
10	(Not 2)	**1**	**11**	N/A
12	1 or 5	N/A	**12**	N/A

As a summary of the conclusions:

- 1 versus 2, 5: The comparison of *Debaryomyces hansenii* located with other species with a CUG serine (arrangement 1) or back in its earlier position closer to *S. cerevisiae* (arrangements 2 and 5) had equivocal results, at least with regard to arrangements other than 2 (which was ruled out, but probably more due to the other alterations). Future work to resolve this uncertainty further may be indicated - especially prior to another attempt at creating Ascomycota models (see "Future work", on page 356).
- 1 versus 2 versus 3: The comparison of positions of *D. discoideum* relative to fungi/metazoa and *E. histolytica* indicated that D. discoideum was closer (arrangement 1) to fungi/metazoa than was E. histolytica. This was, however, contradicted by later results (see "Tree search with Non-Fungi/Metazoa Eukaryota", on page 313), which indicated that *D. discoideum* and *E. histolytica* should be together. As noted previously,

[466] It had originally been thought that results from this tree run indicated phylogeny 1 as more likely. This has been determined to be due to a copying error, suggesting like other such occurrences the value of further automation of this process (see "Future work", on page 334).

further work on these, with more proteins and possibly species used (e.g., the addition of actin and *Hartmannella cantabrigiensis*) is suggested.

- 1 versus 11 versus 12: This was to try to determine the positions of Viridiplantae+Kinetoplastida and Alveolata vis-à-vis Fungi/Metazoa. The results were contradictory between subsets. Tree search runs (see "Tree search with Eukaryota (subset)", on page 300, and "Tree search with Non-Fungi/Metazoa Eukaryota", on page 313) were consequently conducted.
- 1 versus 9 versus 10: This was to determine the position of Cetartiodactyla vis-à-vis Fungi/Metazoa. A later tree search (see "Tree search with Mammalia (subset)", on page 316) indicated that no hypothesis tested was correct.

In hindsight, this round of tree rearrangements was not very successful overall, other than in helping to develop some programmatic aspects (e.g., better subset creation) of the process. This lack of success was at least partially due to the limits of the existing tree rearrangement process - see "Future work", on page 334.

Tree searches

It was concluded that:

- the number of possible rearrangements was too many to be done via the current (partially) manual method (see "Future work", on page 334); and
- the understanding of adjustments to better do tree analysis (e.g., improved species subset creation - see "Species subsets", on page 101) should enable

some tree searches to be conducted in a reasonable amount of time (albeit still more time than for tree rearrangement runs).

Thus, some tree searches (in which MrBayes was allowed to attempt various random rearrangements of a reasonable starting tree) were done. (Please note that the distances for the "phylogram" forms of the tree searches are from that tree search, not from the final tree, and are more uncertain due to less runs/data being used to determine them.)

Tree search with Eukaryota (subset)

Due to the uncertainties regarding the relationship of Alveolata, Kinetoplastida, Viridiplantae, and other eukaryotes, a tree search was run concentrating on these species. This search was done using 2 runs with 4 chains each, 30000 generations (3000 samples), and a burnin for sumt (tree extraction) processing of 2425. The dataset had 6098 amino acids from 19 proteins (counting ADH1 Alpha/Beta/Gamma as 1 protein). The results[467] are shown in Figure 4.T.s.eukaryota.p (on page 301) and Figure 4.T.s.eukaryota.c (on page 302).

[467] The numbers at various nodes (e.g., "1.0") are indications of with what proportion this tree arrangement was seen among the trees examined after the "burnin" period. Note that the tree is shown with "full" species (with a ".#" (".number") after the species name, e.g., "*Arabidopsis_thaliana.1*", "*Arabidopsis_thaliana.2*", "*Homo_sapiens.01*", or "*Homo_sapiens.02*"), and their groupings' consistent "1.0" results for proportion seen is artificial (see 'Creation of "full" species', on page 68). (*Homo sapiens*' numbers below 10 are prefixed with a 0 due to the highest number for *Homo sapiens* going above 9 - otherwise, "*Homo_sapiens.2*" would be sorted after "*Homo_sapiens.10*". The situation is similar for *P. carinii*, mainly due to multiple DHFR sequences.)

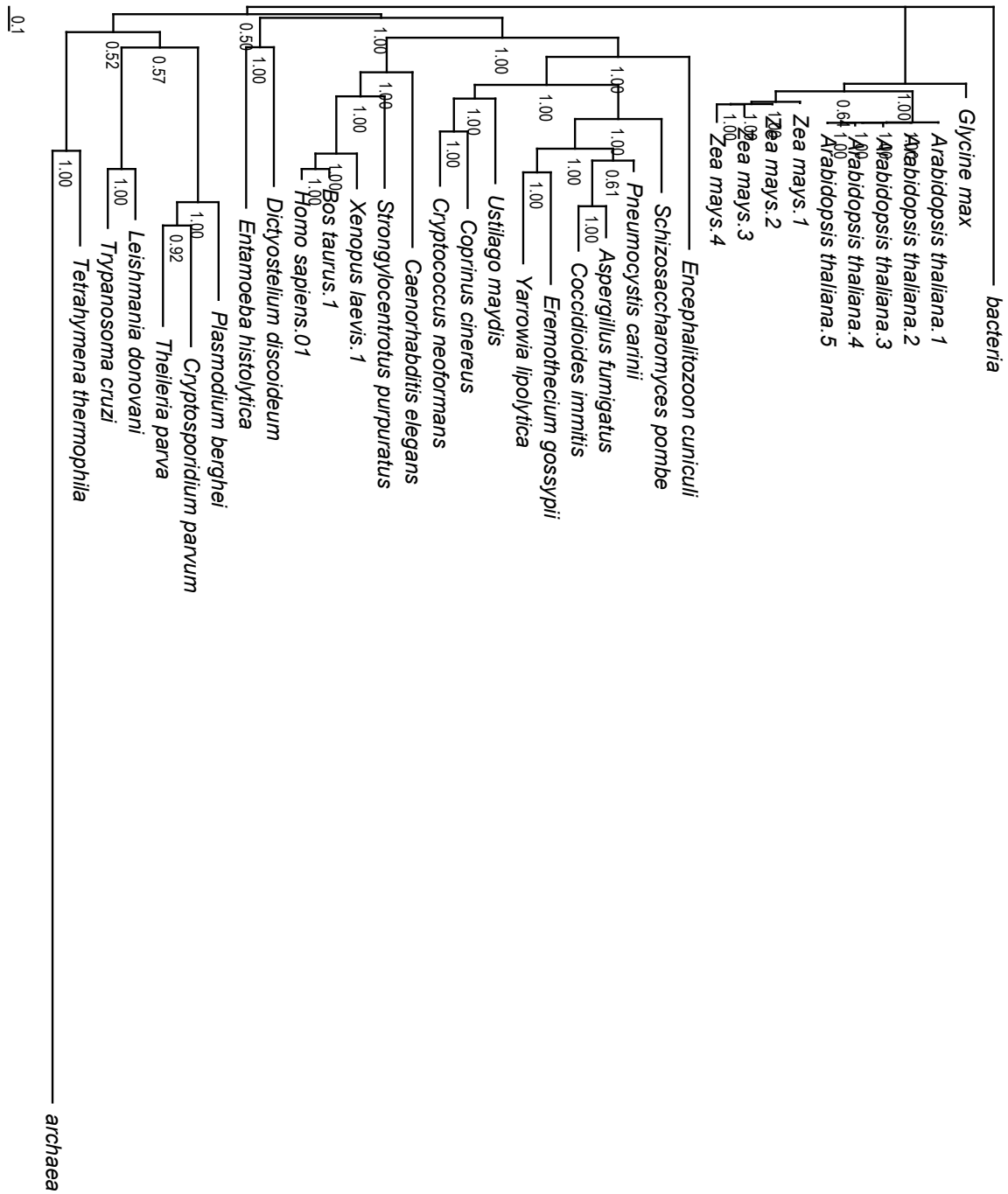


Figure 4.T.s.eukaryota.p: Tree search of Eukaryota (subset), phylogram

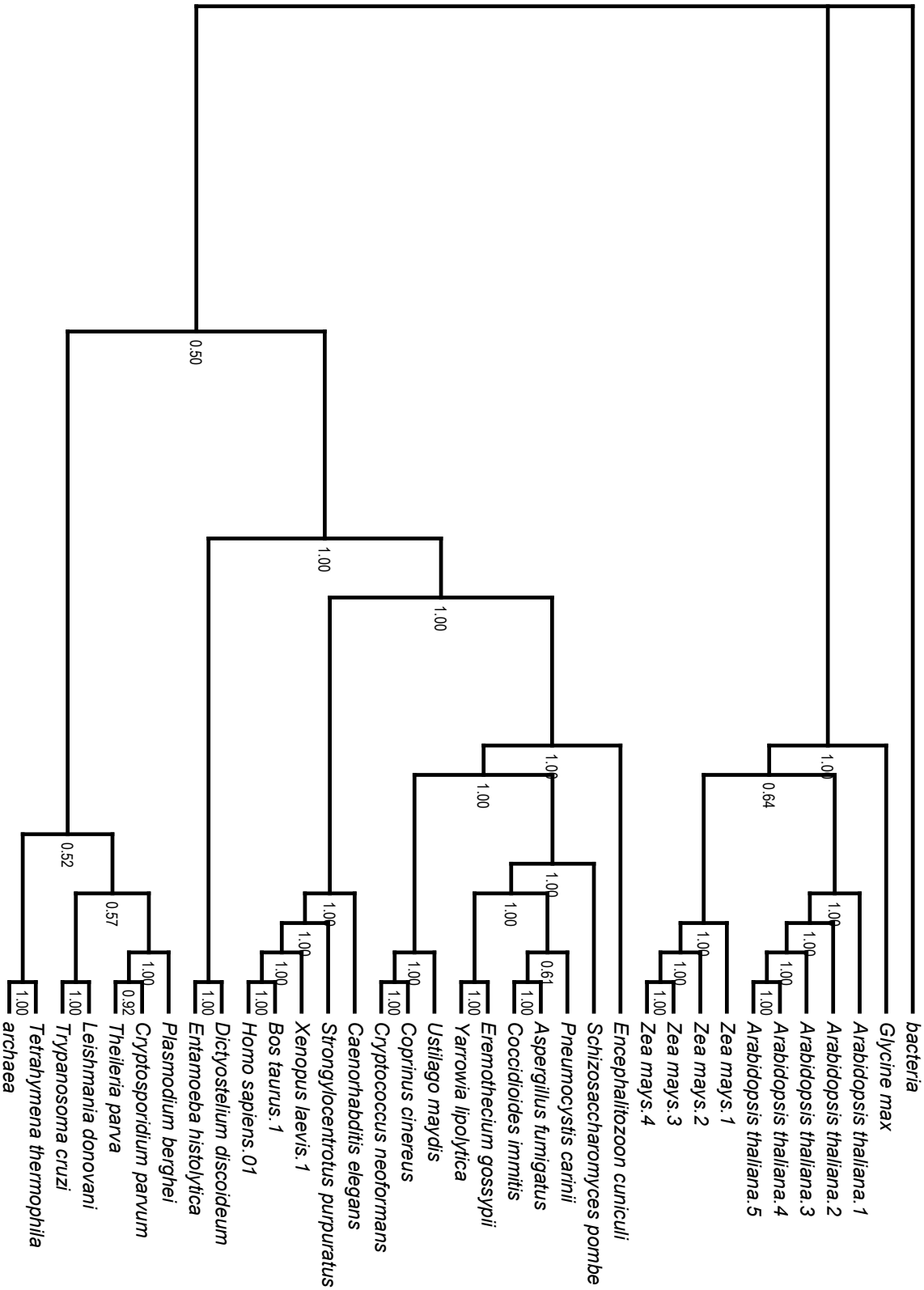


Figure 4.T.s.eukaryota.c: Tree search of Eukaryota (subset), cladogram

The above led to the conclusion that Viridiplantae were basal, with Alveolata and Kinetoplastida then branching off in one clade and Fungi/Metazoa[468] in the other. As well as some node support values being lower than would be desired, it was noted that there appeared to be long branch attraction (see footnote 52 under "Tree construction methods", on page 27) taking place between Archaea[469] and the likewise long-branching *Tetrahymena thermophila*. After the use of better group sequence creation techniques (see "Further sequence processing: Group sequence creation", on page 96) was unsuccessful at eliminating this problem (see "Tree search with Insecta, some other Eukaryota", on page 309), Archaea[470] were eliminated (unless specified otherwise), even as an outgroup sequence, for tree work with Eukaryota.

---

[468] "Fungi/Metazoa" in the above did include possible fungi/metazoa. Please note that this tree unfortunately did use a constraint for *Dictyostelium discoideum* and *Entamoeba histolytica* being together with fungi/metazoa. In hindsight, this was an error, but tree searches had been sufficiently problematic that some assistance was felt necessary, and no usable DHFR sequence is known for the two species in question, making them lower priority. A partial rerun of this tree search, without the problematic Archaea and with the addition of, for instance, actin (see footnote 463 under "Subset 10", on page 267), is advisable. The addition of further, non-parasitic species (see footnote 235 under "Tree distances", on page 115) from the Alveolata and Kinetoplastida would also be of assistance.

[469] Please note the length of some of the branches in the tree, including even among Eukaryota alone (e.g., see Figure 4.T.nfm, on page 328, noting the mutations per site scale of 0.1 as compared to the branch lengths). Distances between kingdoms were unsurprisingly larger - above 1 in general, meaning an expectation that at least one mutation will have happened per site. Some level of long branch attraction is therefore (highly) unsurprising. Indeed, it appears likely that other methods of tree building would have had worse problems (Anderson & Swofford 2004; Ranwez & Gascuel 2001).

[470] The position of Archaea may be especially sensitive to bias due to only one protein, RecA/RadA, being known for many species in the dataset; however, this protein is adequate for work within the Archaea (see "Initial sources", on page 72). Further investigation of whether, for instance, this implies correlated mutations in RecA/RadA is of interest. (The usage of rRNA , tRNA, *etc.* in Archaea may be of interest for comparative purposes.) Another possibility is because of the branch lengths within Archaea, some of which were above 1. On the other hand, upon examination of the current Archaeal tree ("archaea.phy"), the problem may be human error (at least partially due to time pressure), namely not correctly rooting the tree, as intended, between Crenarchaeota and Euryarchaeota and the subsequent addition of *Archaeoglobus fulgiditus* (see "Initial sources", on page 72) in a consequently further incorrect location. A rerun of the Eukaryota and (see below) Insecta searches with this error corrected (following tree runs for

It was also noted that:

- unlike with most prior runs (data not shown) - chain swaps[471] were taking place (68%-90% successful attempts), but
- these swaps were probably successful only because the "temperature" differences between the chains were minimal (cold chain 1.0, hottest chain 0.97), making them unlikely to be useful.

See for further information:

- the following footnotes under “MrBayes code alterations”:
  - 197 on page 98
  - 202 on page 100
- Appendix J: MrBayes review/explanation, on page 379

Tree search with Proteobacteria (subset)

Please note that this and following tree work used the improved grouping technique using distances (see “Further sequence processing: Group sequence creation”, on page 96). Also note that the covarion option (see footnote 200 under “MrBayes code alterations”, on page 99) was not used at or after this point. Given the number of possible rearrangements of Bacteria, and that this research focused on Eukaryota, it was decided that while a tree search would be run on

---

better distances for Archaea to derive a reasonable outgroup sequence) may be of interest.

[471] Each “chain swap” results in the effective transfer of information from one chain (e.g., one freer to vary - at a higher “temperature” in the normal sense) to another (less free to vary, or “colder”), and vice-versa. In other words, it is how possibilities explored by the “looser” chains are communicated to the main chain, if they appear close enough for this to make sense.

Proteobacteria, no further rearrangements on Bacteria[472] would be done. This search was with 2 runs, 300000 generations (3000 samples), and a burnin for sumt of 2925. The dataset had 2850 amino acids in 7 proteins. (This admittedly used fewer samples remaining after burnin than would be desirable - an effort was made to have at least 100 in most cases - but, again, the concentration of this research was on Eukaryota.) The results are shown in the trees on pages 306-308.

[472] Since no bacterial DHFRs were used (or intended to be used, unlike - initially - bacterial TS), the primary reason for this decision was for the use of bacteria for an outgroup sequence - a process that uses the phylogeny of said outgroup (see “Further sequence processing: Group sequence creation”, on page 96). A secondary purpose for this was that the bacterial tree had earlier been useful in detecting errors in methodology (see footnote 53 under “Tree construction methods”, on page 30).

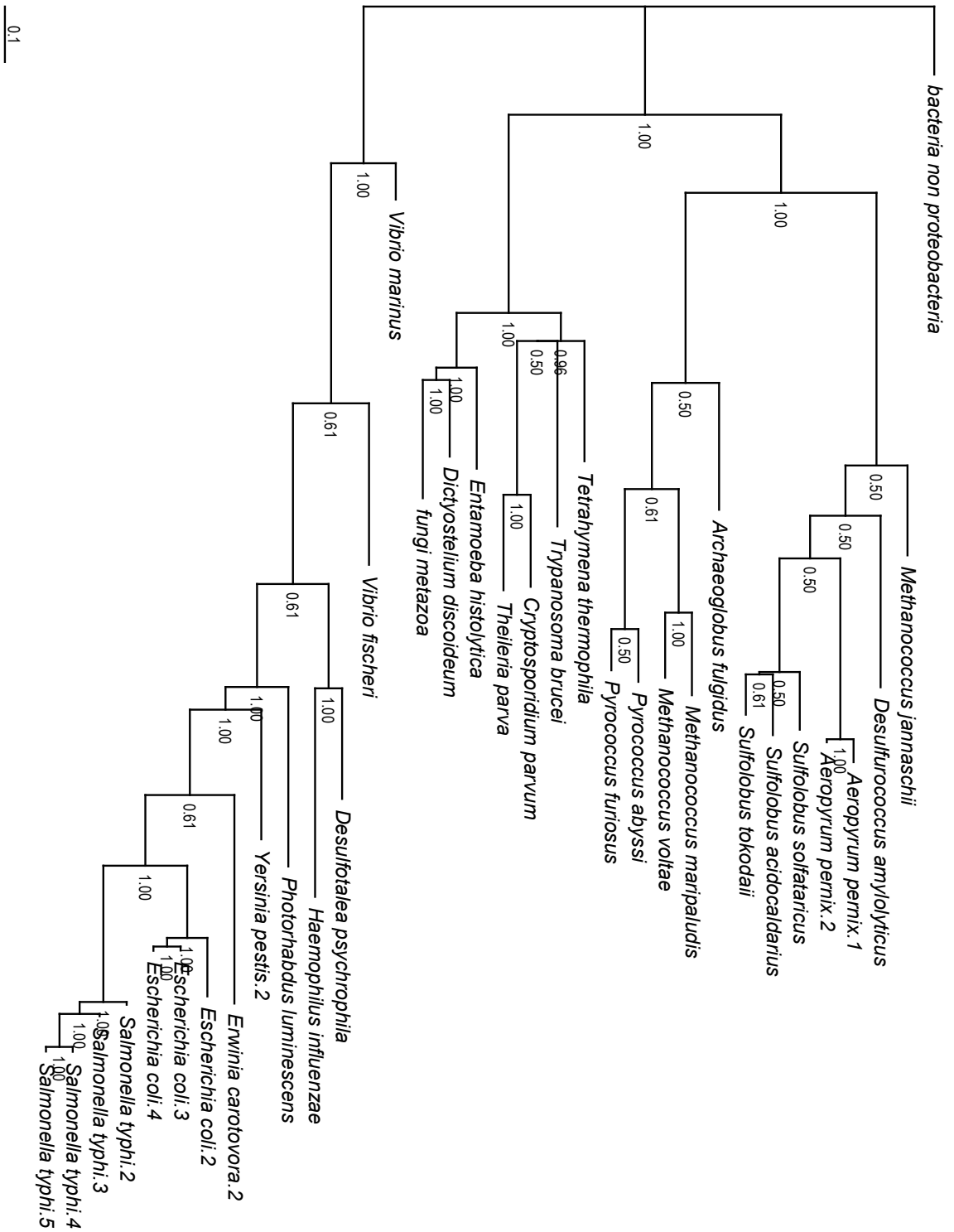


Figure 4.T.s.proteobact.p: Tree search of Proteobacteria (subset), plus outgroups; phylogram

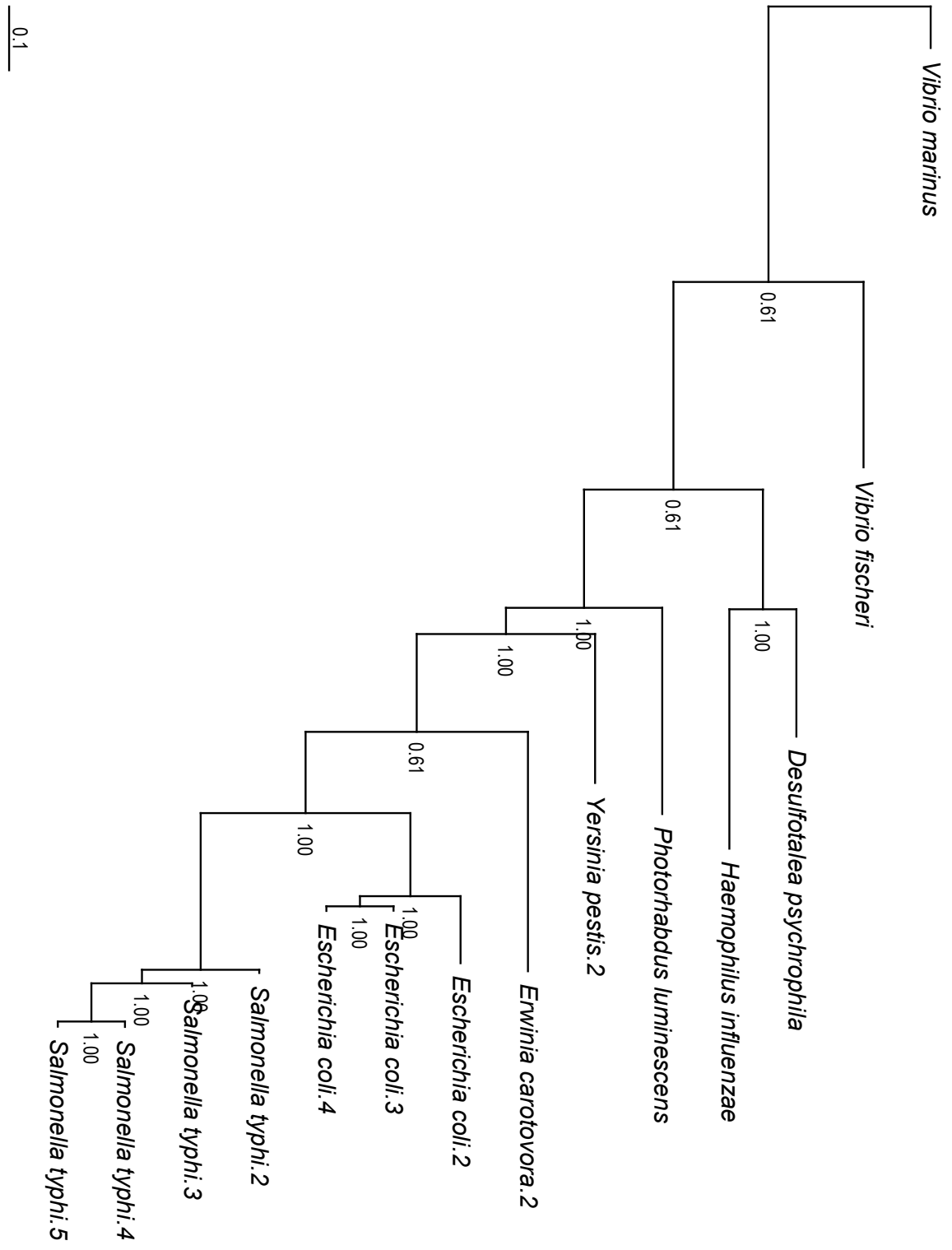


Figure 4.T.s.proteobact.p.proteobact: Tree search of Proteobacteria plus outgroups, Proteobacteria only shown, phylogram

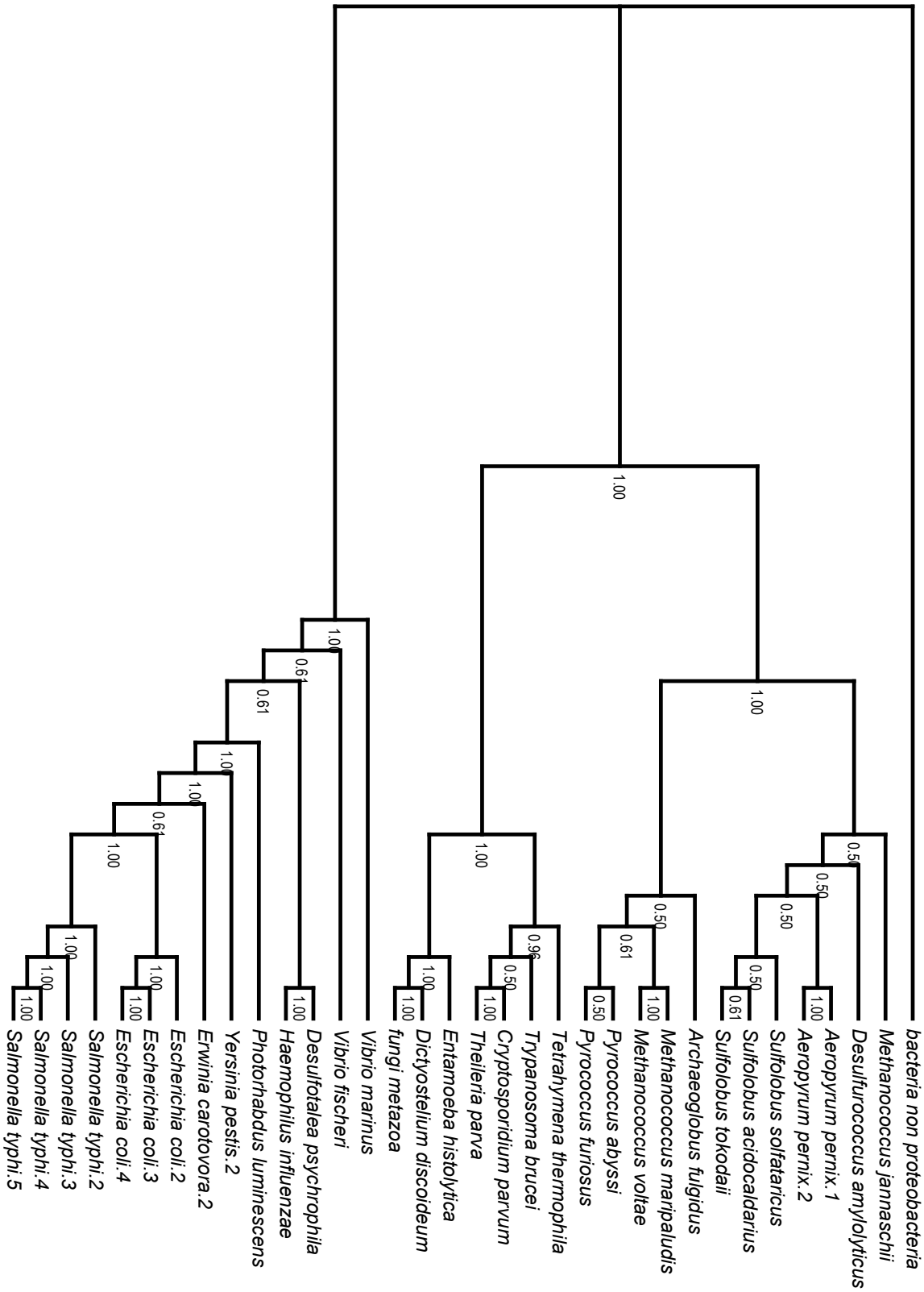


Figure 4.T.s.proteobact.c Tree search of Proteobacteria (subset), plus outgroups; cladogram

Tree search with Insecta, some other Eukaryota

Given the inconclusive nature of the above search regarding Insecta, including of species with DHFR sequences, and that some copying errors had been made during the tree rearrangements, preventing a valid evaluation of alternate Insecta arrangements[473], a tree search run was performed. (Please note that SA and Adapt were used for this and subsequent tree work.) This tree search used 2 runs, 300000 generations (3000 samples), and a burnin for sumt of 2250; the dataset had 5668 amino acids in 18 proteins (with ADH1 counted as 1 protein). The results are shown in the figures on pages 310-312.

[473] Admittedly, the Insecta evaluation might have been disrupted in any event by the usage of the original Ecdysozoa and Lophotrochozoa assumption (see “First round of tree rearrangements”, on page 203).

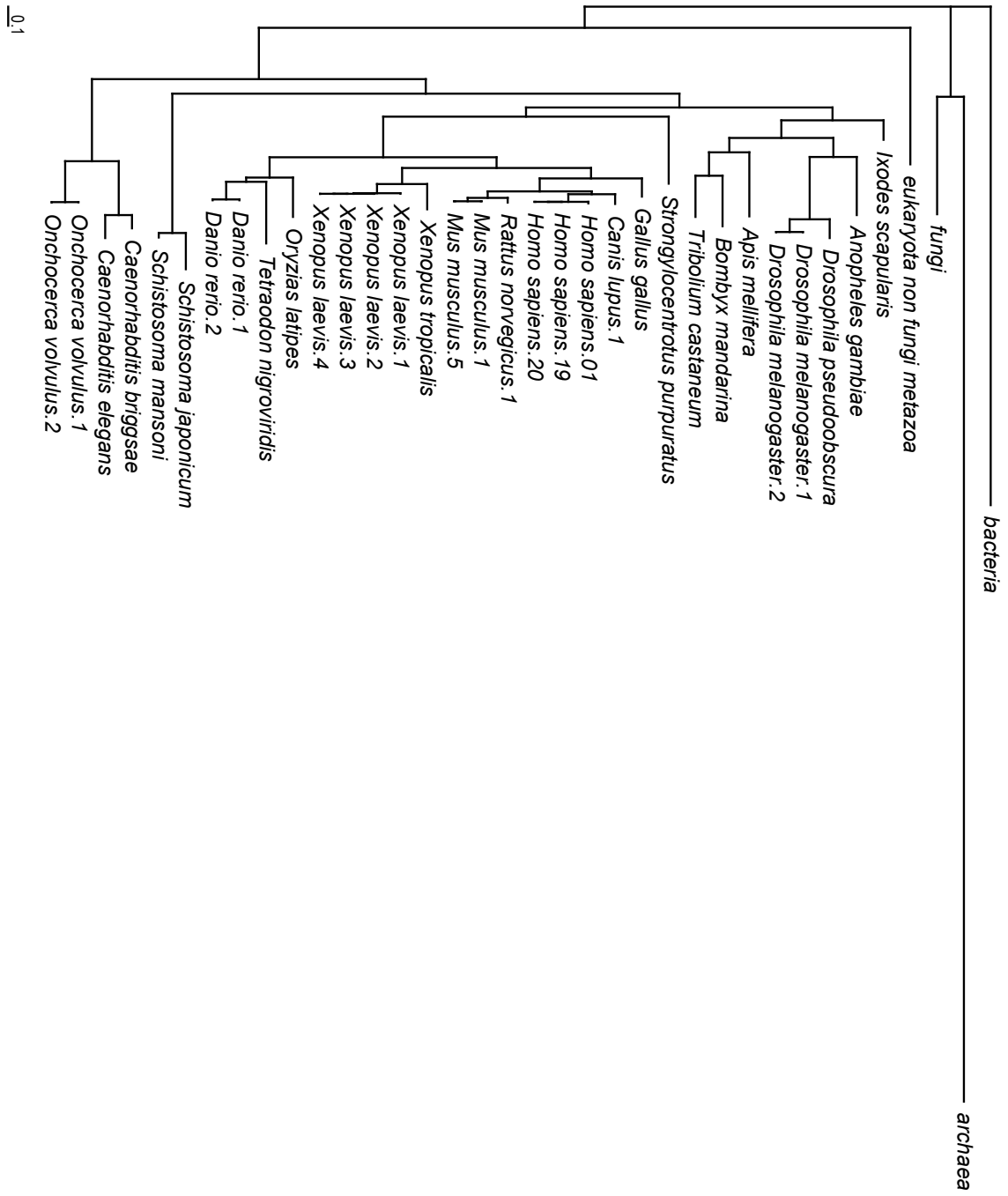


Figure 4.T.s.insecta.p: Tree search of Insecta (and others), phylogram

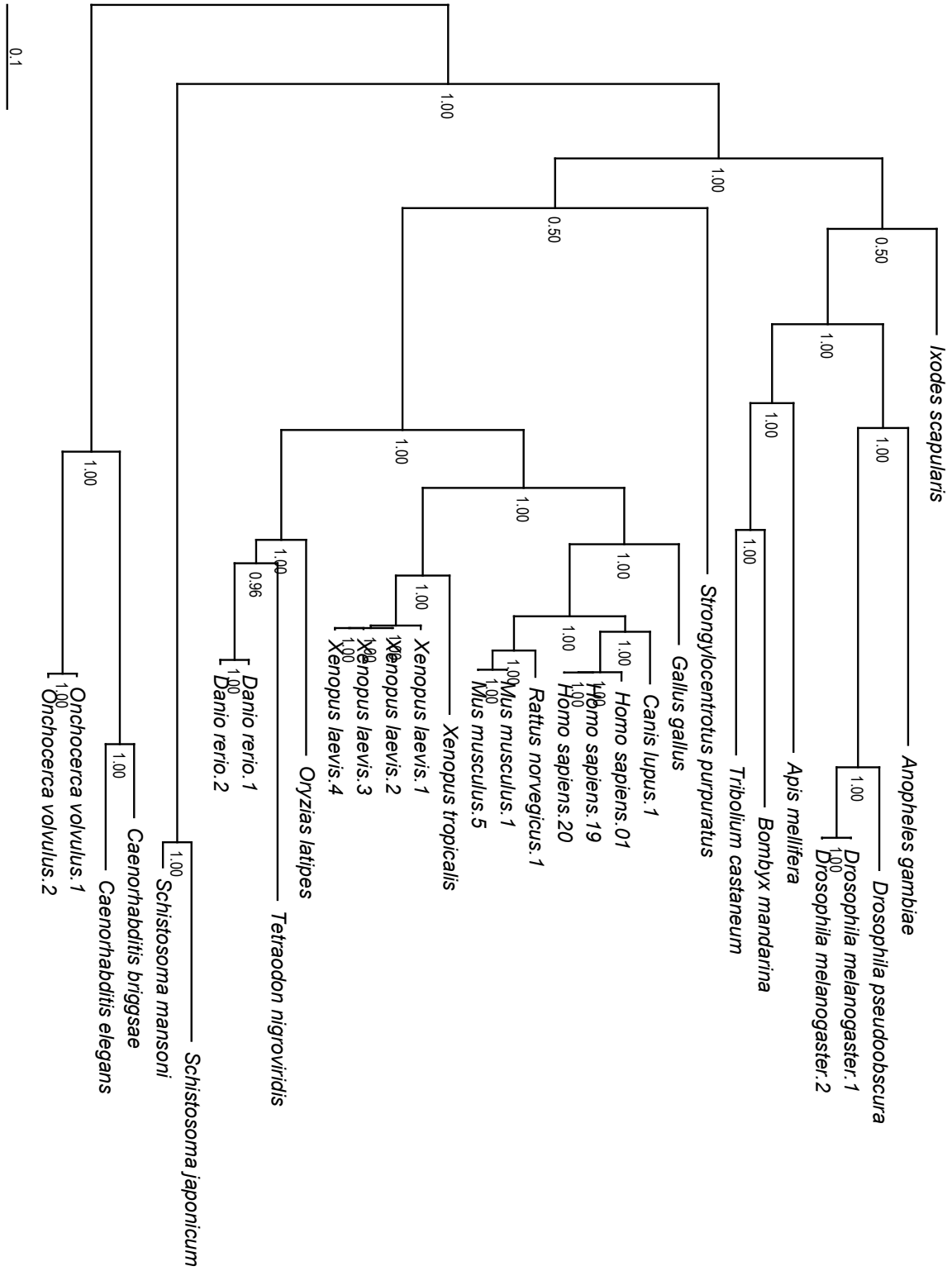


Figure 4.T.s.insecta.p.metazoa: Tree search of Insecta (and others), Metazoa only shown, phylogram.

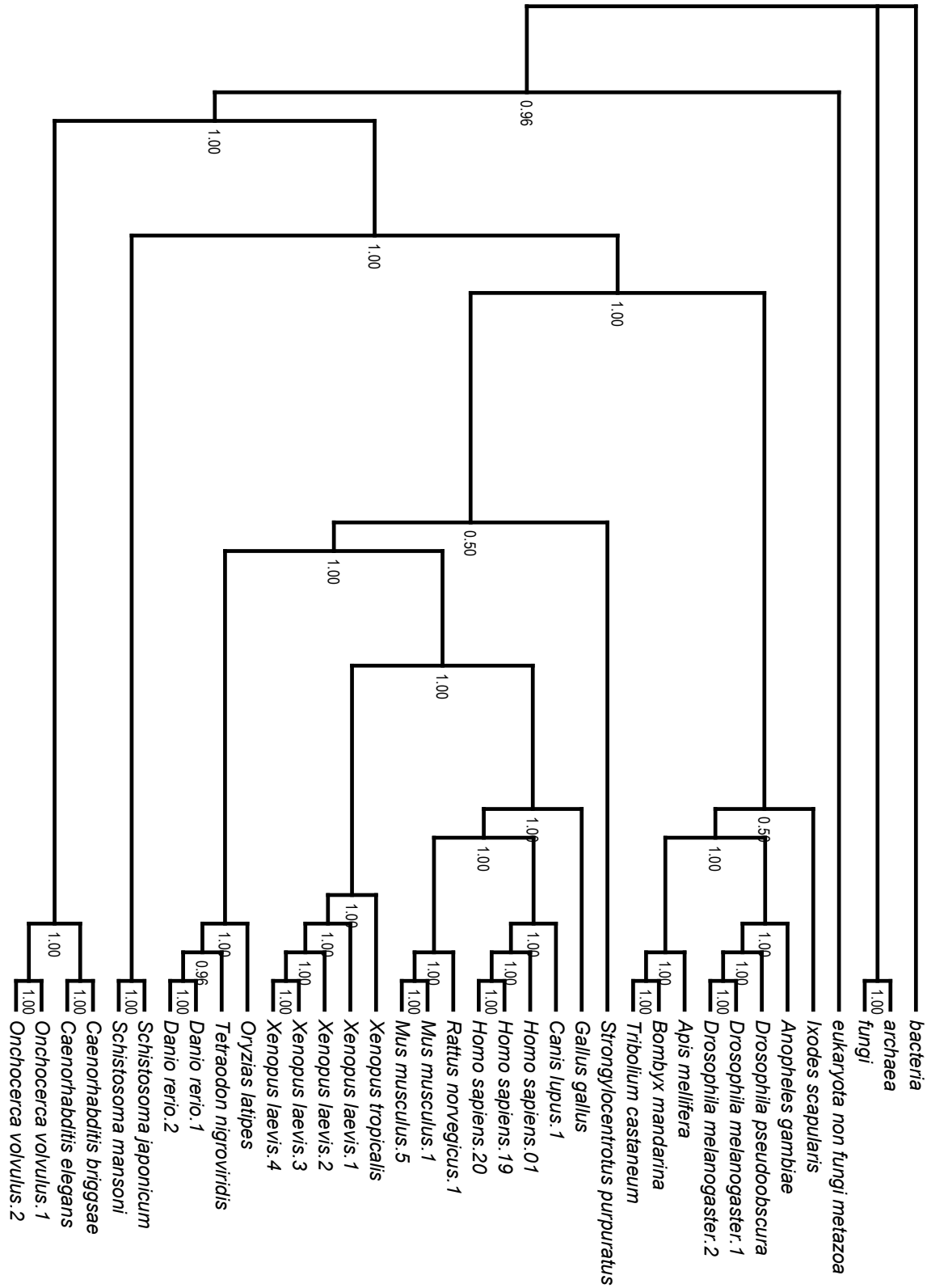


Figure 4.T.s.insecta.c: Tree search of Insecta (and others), cladogram

Except for the uncertainty in the positioning of *Ixodes scapularis* (the blacklegged tick)[474], and probable long-branch attraction between the "fungi" and (non-cladal) "eukaryota non-fungi/metazoa" group sequences, the results appear to be well supported. As well as the arrangement inside Protostomia, the support for tree arrangement 4 from the first round (see "Summary of first round results", on page 264) is notable.

Tree search with Non-Fungi/Metazoa Eukaryota

Due to the problems encountered earlier with non-fungi/metazoa placements, and the importance of several *Plasmodium* and *Cryptosporidium* species as having known, usable DHFR[475] structures, a tree search was done focusing on non-fungi/metazoa. This search used 2 runs with 400,000 generations (4000 samples) each, with a burnin of 3000, using a dataset of 3627 amino acids among 11 proteins. The tree results are shown on pages 314-315.

---

[474] In the alternative tree, it was in a clade with *Strongylocentrotus purpuratus* (the purple sea urchin). Some uncertainty in the location of this species is not surprising, insofar as it only has two proteins in the database (CuZnSOD and UBC, neither of which are present for *Strongylocentrotus purpuratus;* this may point to a remaining problem with missing data). A mild degree of long branch attraction (see footnote 52 under "Tree construction methods", on page 27) may also be taking place, given that *Ixodes scapularis* is the only non-Endopterygota present among the Protostomia present.

[475] To be precise, DHFR/TS.

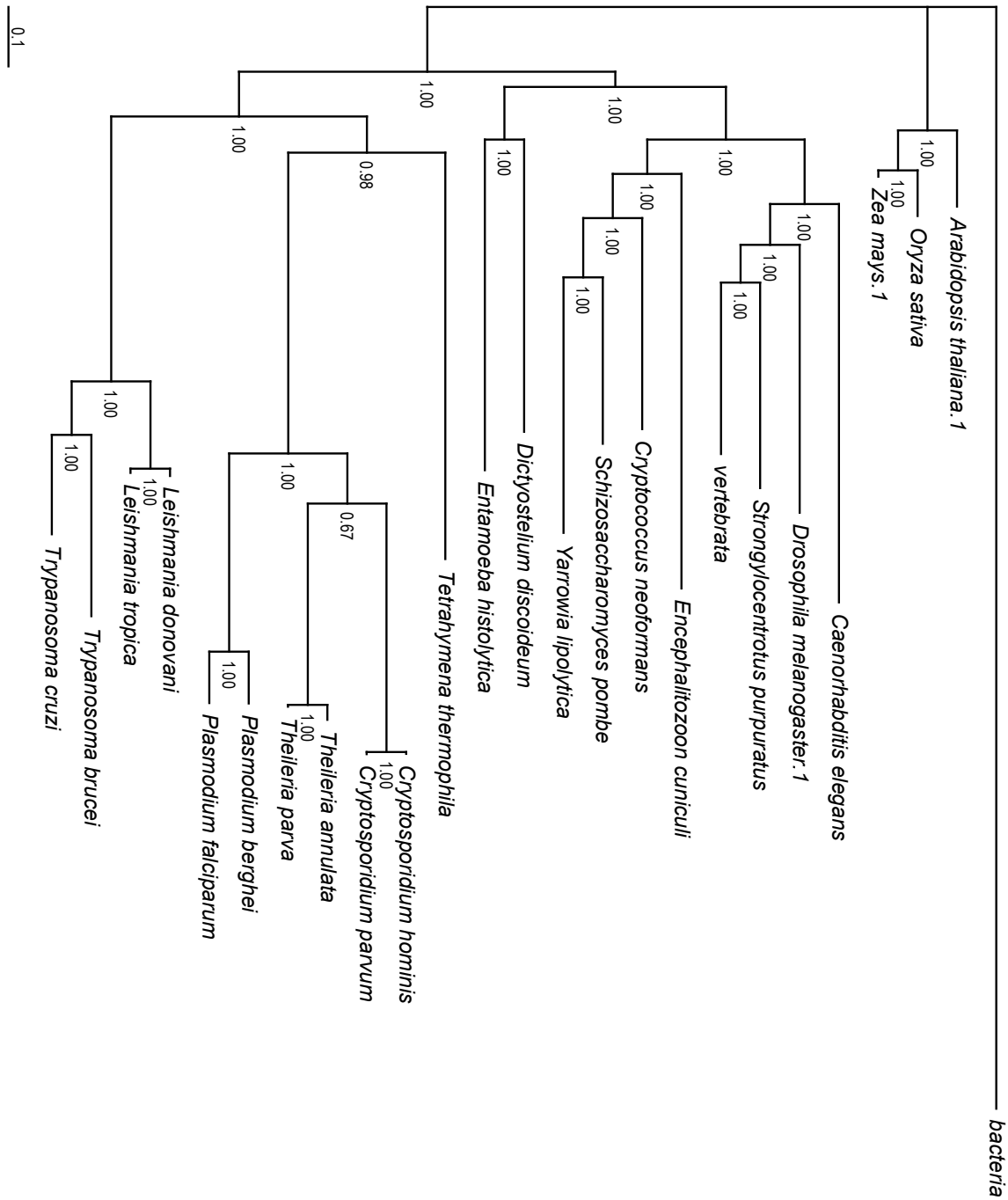


Figure 4.T.s.nfm.p: Tree search of Non-Fungi/Metazoa Eukaryota, plus outgroups; phylogram

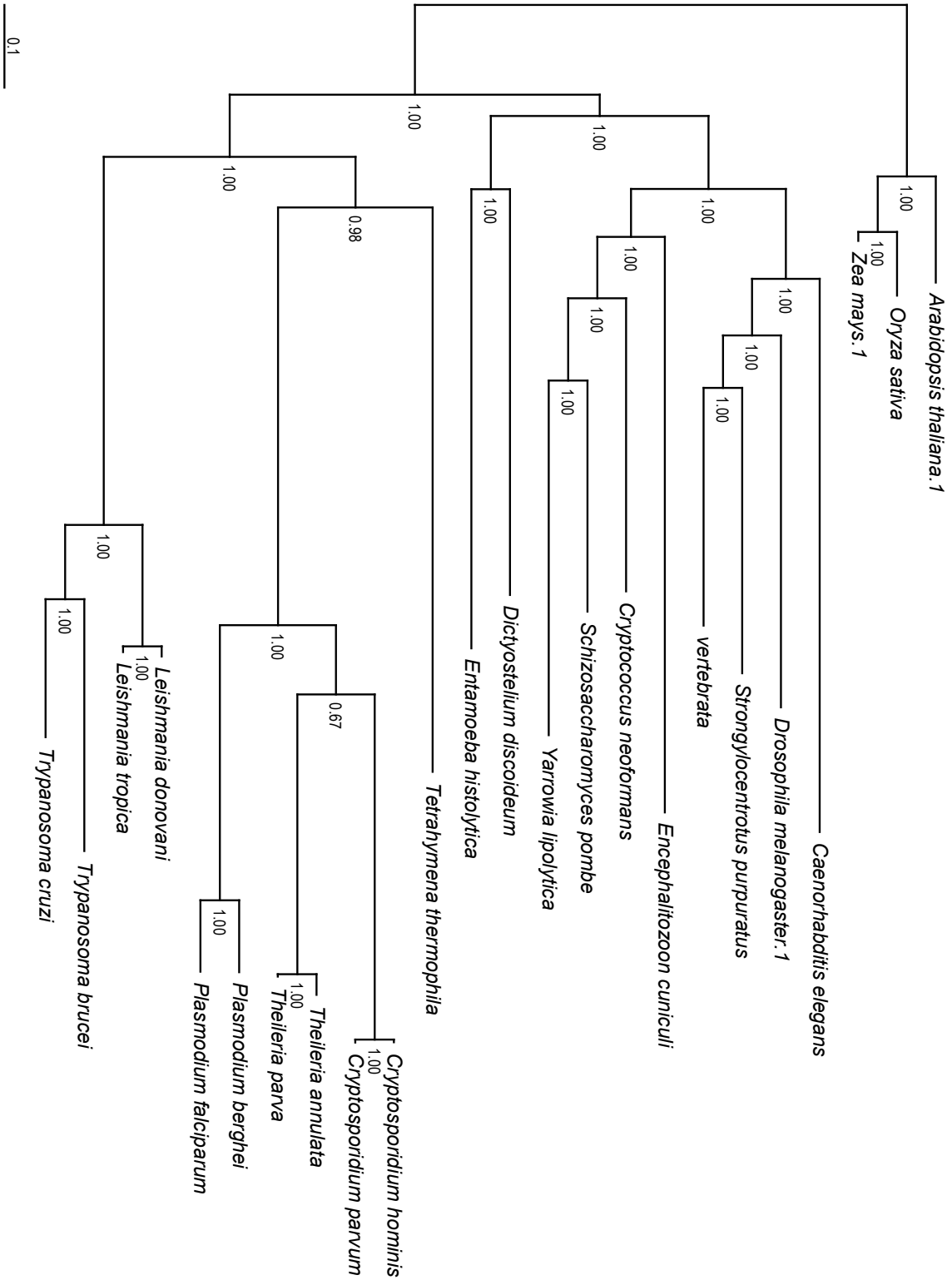


Figure 4.T.s.nfm.p.eukaryota: Tree search for Non-Fungi/Metazoa Eukaryota (and others), Eukaryota only shown, phylogram

Tree search with Mammalia (subset)

Please note that (at least some) DHFR sequences were included in this and later tree work. Given the:

- confusion regarding the proper positioning of various mammal species noted above and in other work (Kullberg *et al.* 2006); and
- importance of *Homo sapiens* and *Mus musculus* (with their DHFR structures used as templates) for the present work,

a tree search was run on Mammalia[476] and some other Tetrapoda (chiefly those with ADH1, Hemoglobin V/Alpha, and/or Myoglobin sequences available), plus group sequences including from other Vertebrata (generally fish). The search was done using 6 runs in parallel, 400000 generations (4000 samples), with a burnin for sumt of 3501. There were 5411 amino acids in 18 proteins (with ADH1 as 1) used. The results are shown in pages 317-319.

[476] The focus among Mammalia was on Primates and Rodentia with known DHFR sequences, although Cetartiodactyla and Carnivora with known DHFR sequences were also deliberately included. Note that opossums (*Didelphis marsupialis* and *Monodelphis domestica*), *Oryctolagus cuniculus* (rabbit - a member of a species group (Lagomorphs) the position of which is in dispute (Easteal 1990; Kullberg *et al.* 2006), like other groups such as primates and Rodentia), and *Cebus apella* (Capuchin monkey) were specifically removed from the subset used, given earlier problems with them as detected by "compare.trees.problems.pl". This elimination may be at least partially responsible for the problematic distance for, for instance, *Oryctolagus cuniculus* in the final tree - in other words, that there are fewer tree runs using *Oryctolagus cuniculus* may be the cause for its distance in the final tree being extremely short, as noted during the defense of this dissertation work. (The distance between the divergence of Lagomorpha and Rodentia and their divergence from other placental mammals was so short as to not be visible on the printout of the Eukaryota portion of the final tree.) However, given that the inclusion of *Oryctolagus cuniculus* and *Cebus apella* in prior tree searches yielded a result (among others) of primates as non-cladal, and neither has usable/known DHFR sequences, this was not felt to be worrisomely problematic. Exactly why the inclusion of Lagomorphs is problematic is a question for further research; the problems with *Cebus apella* can be attributed to lack of sequences (only 3 - CuZnSOD, Hemoglobin Alpha/V, and Myoglobin - in the alignment dataset used), with similar difficulties with opossums.

Figure 4.T.s.mammalia.p: Tree search of Mammalia (subset), plus outgroups; phylogram (see other figures for node support values)

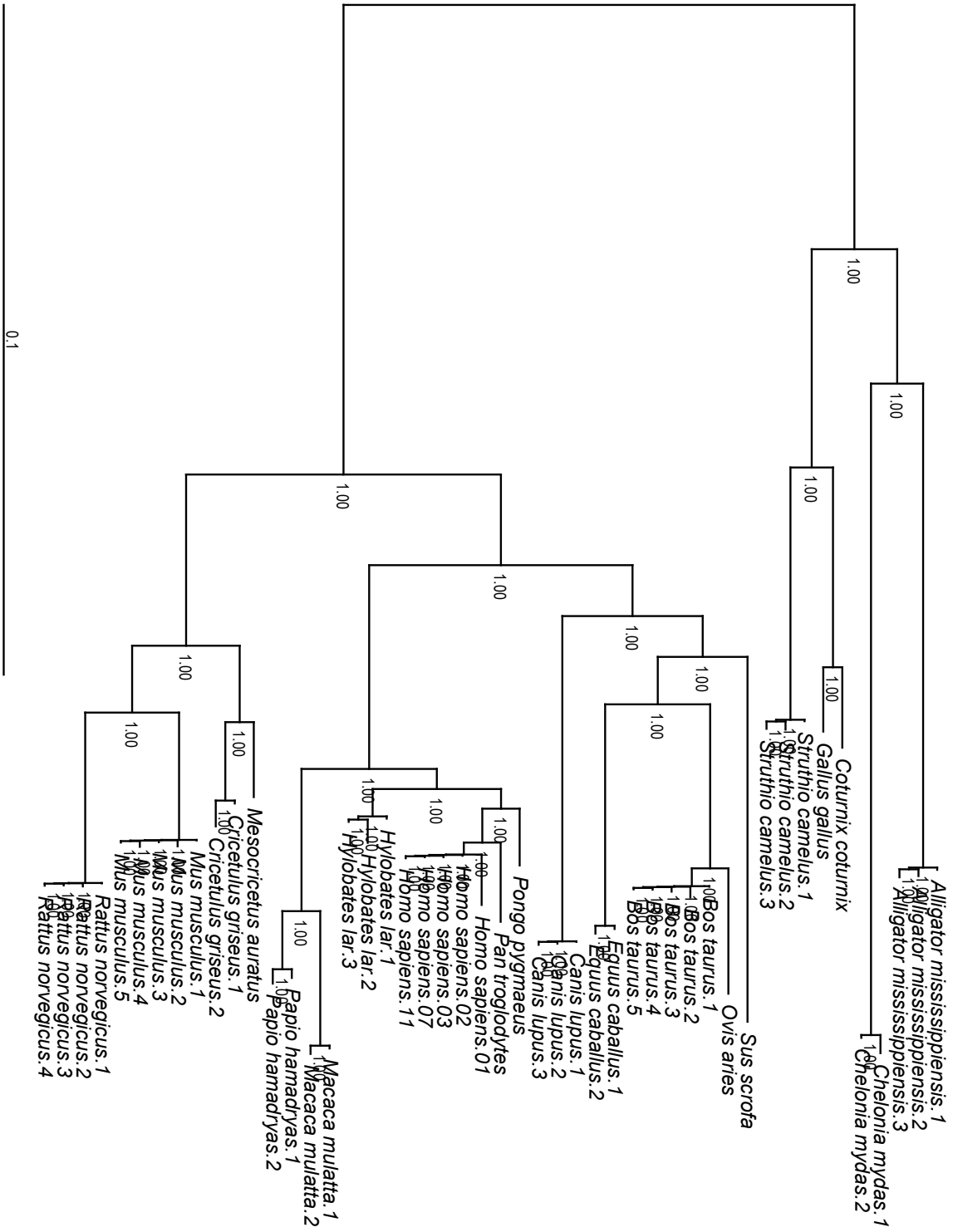


Figure 4.T.s.mammalia.p.tetrapoda: Tree search of Mammalia (subset; plus others), Tetrapoda only shown, phylogram

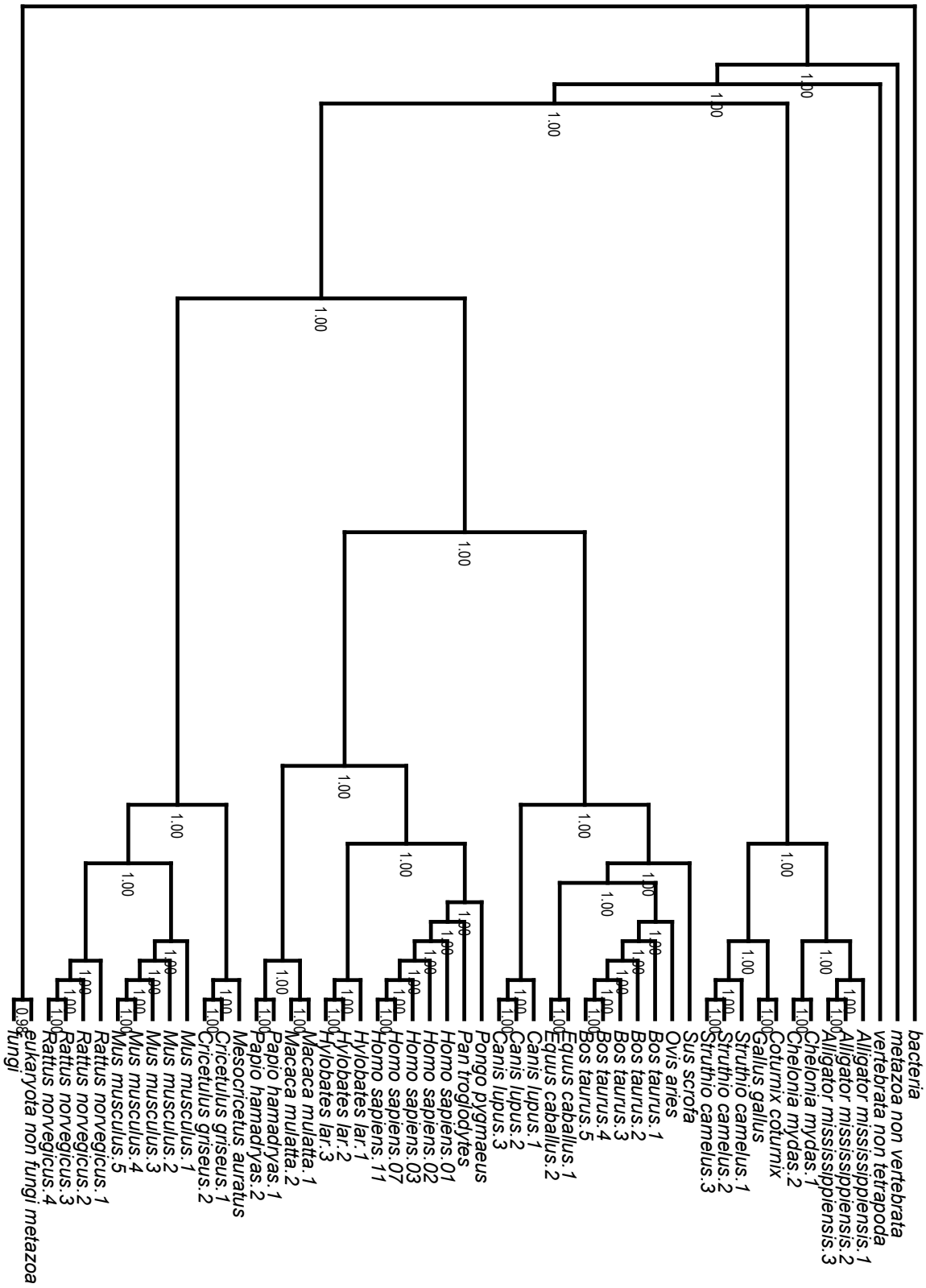


Figure 4.T.s.mammalia.c: Tree search of Mammalia (subset), plus outgroups; cladogram

Note that, for tetrapoda (see Figure 4.T.s.mammalia.p.tetrapoda, on page 318), all support values are 1. On the other hand, some probable long-branch attraction effects (see footnote 52 under “Tree construction methods”, on page 27) were noted between the groups “fungi” and “eukaryota non-fungi/metazoa”, possibly due to the non-cladal nature of the latter[477].

Tree rearrangement for *P. carinii*, *S. pombe*

The proper positioning of *P. carinii* and *S. pombe* was considered:

- questionable (lack of sufficient prior data, for instance; some difficulty/uncertainty had been encountered during the initial tree construction with regard to these species, with some (not entirely successful) manual adjustments in response to notably long branch lengths); and
- important (given the former’s DHFR (target) structure).

An (additional) tree rearrangement, using the Archiascomycetes (Webster, Weber 2007) grouping of *P. carinii* and *S. pombe* (instead of the prior arrangement of *S. pombe* as branching first) was therefore checked. There were 3 runs for each arrangement, using 4180 amino acids (and some additional data from gap characters - see "Gap determination", on page 139) and 13 proteins (including DHFR). The runs were for 300000 generations each (sample size 3000), using a burnin of 2837.

[477] A rerun of this tree search (including starting with several arrangements of Mammalia) without the latter group’s sequence may be of interest, especially prior to any (further) publication focusing on these findings.

Please see pages 322-326 for the trees[478]; the results were as follows:

Phylogeny Tested	Run[479]	Arithmetic Mean	Harmonic Mean
Original (see page 326)	1	-160,515.71	-160,536.64
	2	-160,520.64	-160,546.41
	3	**-160,373.75**	**-160,403.65**
	Overall[480]	-160,374.85	-160,545.31
Archiascomycetes (see pages 322-325)	1	**-160,340.33**	**-160,377.48**
	2	**-160,430.11**	**-160,448.69**
	3	-160,486.45	-160,509.36
	Overall	**-160,341.43**	**-160,508.26**

While the log probability results are unfortunately somewhat equivocal, it was decided that the Archiascomycetes grouping would be used.

[478] Note that, due to this topological change being relatively late in the process, much of the data used for distances (see "Tree distances", on page 113) was from the "original" topology. This is reflected in the extremely small branch lengths involved in *P. carinii* and *S. pombe*'s final arrangement. (The branch lengths under the prior arrangement were also very small (very little distance between the branching of *S. pombe* and that of *P. carinii*).)

[479] Each run number used identical seeds for both trees; e.g., run 1 for the original and the Archiascomycetes trees used random number X, while run 2 used random number Y for both trees.

[480] The "Overall" value is from MrBayes (the "TOTAL" from "sump") and is a mean of the samples from all three runs (after removing burnin).

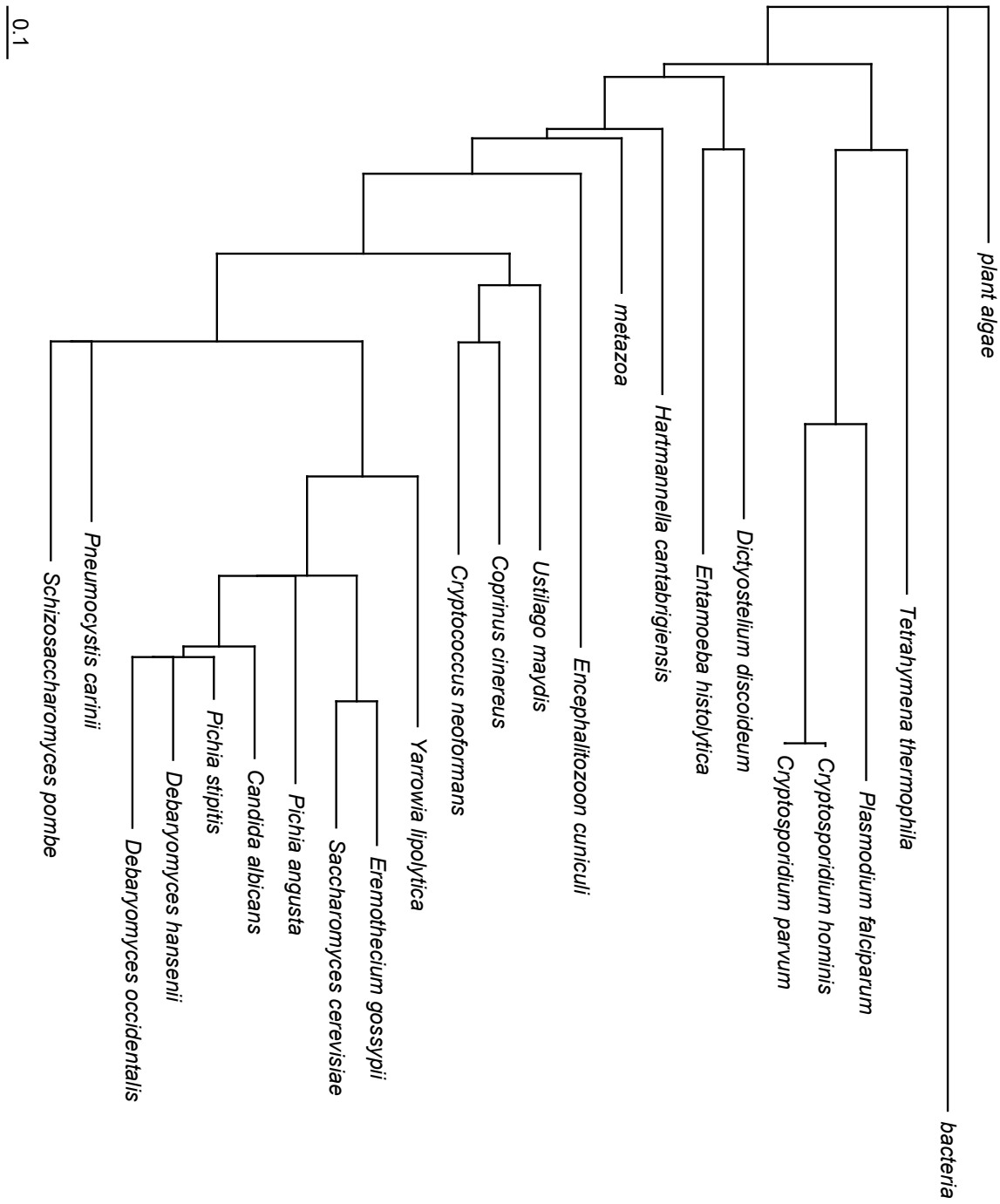


Figure 4.T.r7.s15.c.p: Round 7 subset 15 of final tree, phylogram

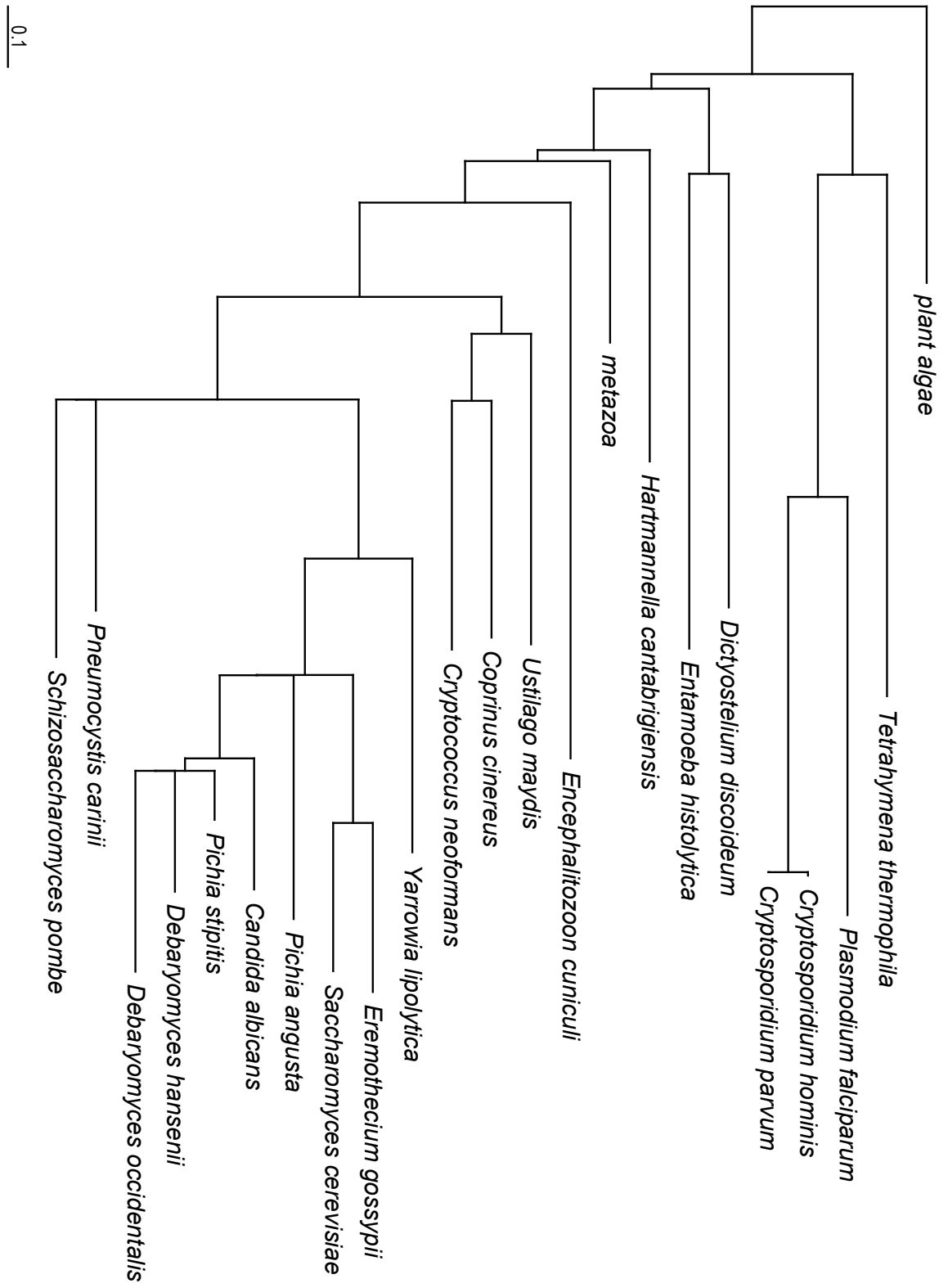


Figure 4.T.r7.s15.c.p.eukaryota: Round 7 subset 15 of final tree, Eukaryota only shown, phylogram

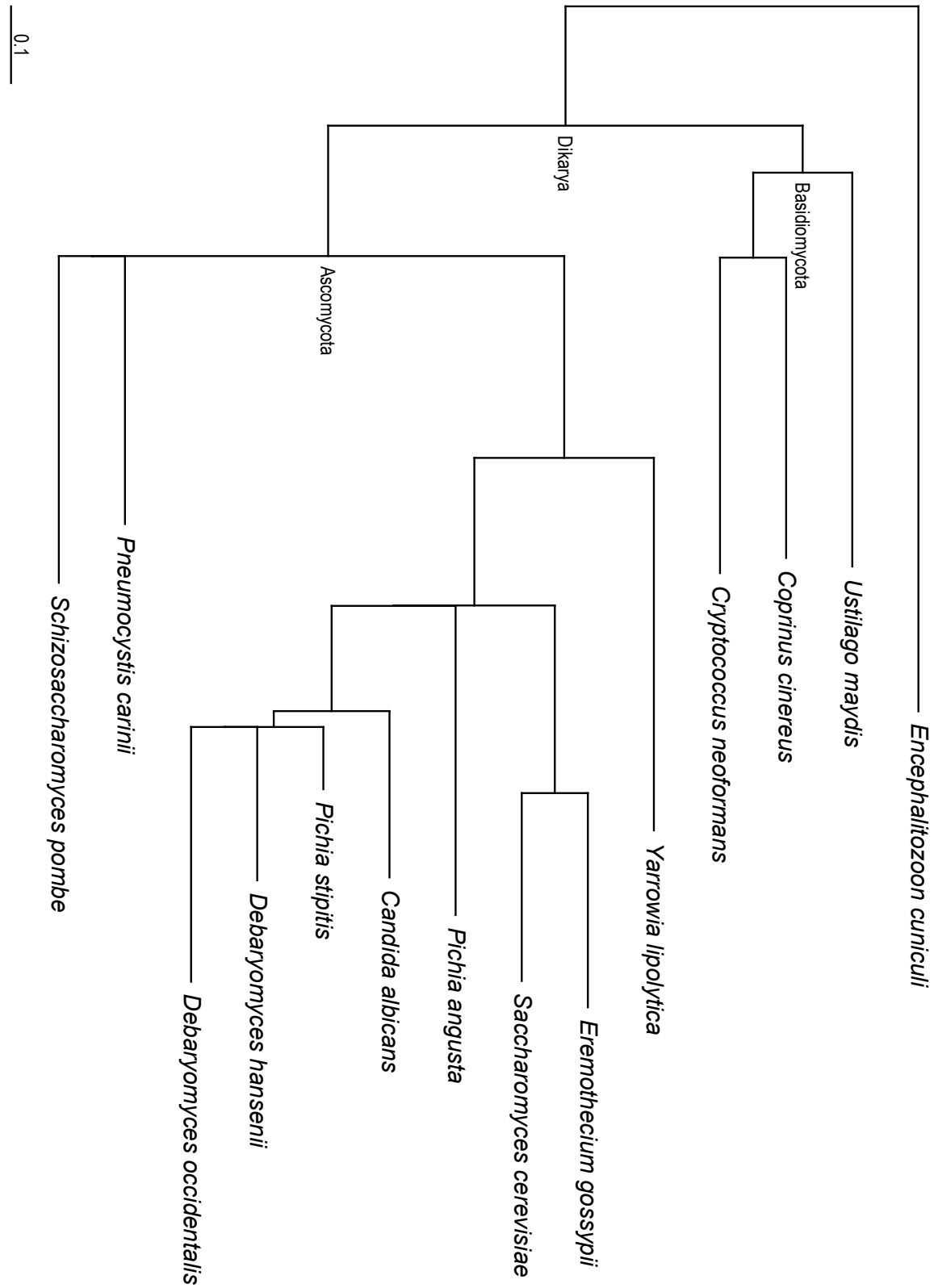


Figure 4.T.r7.s15.c.p.fungi: Round 7 subset 15 of final tree, Fungi only shown, phylogram

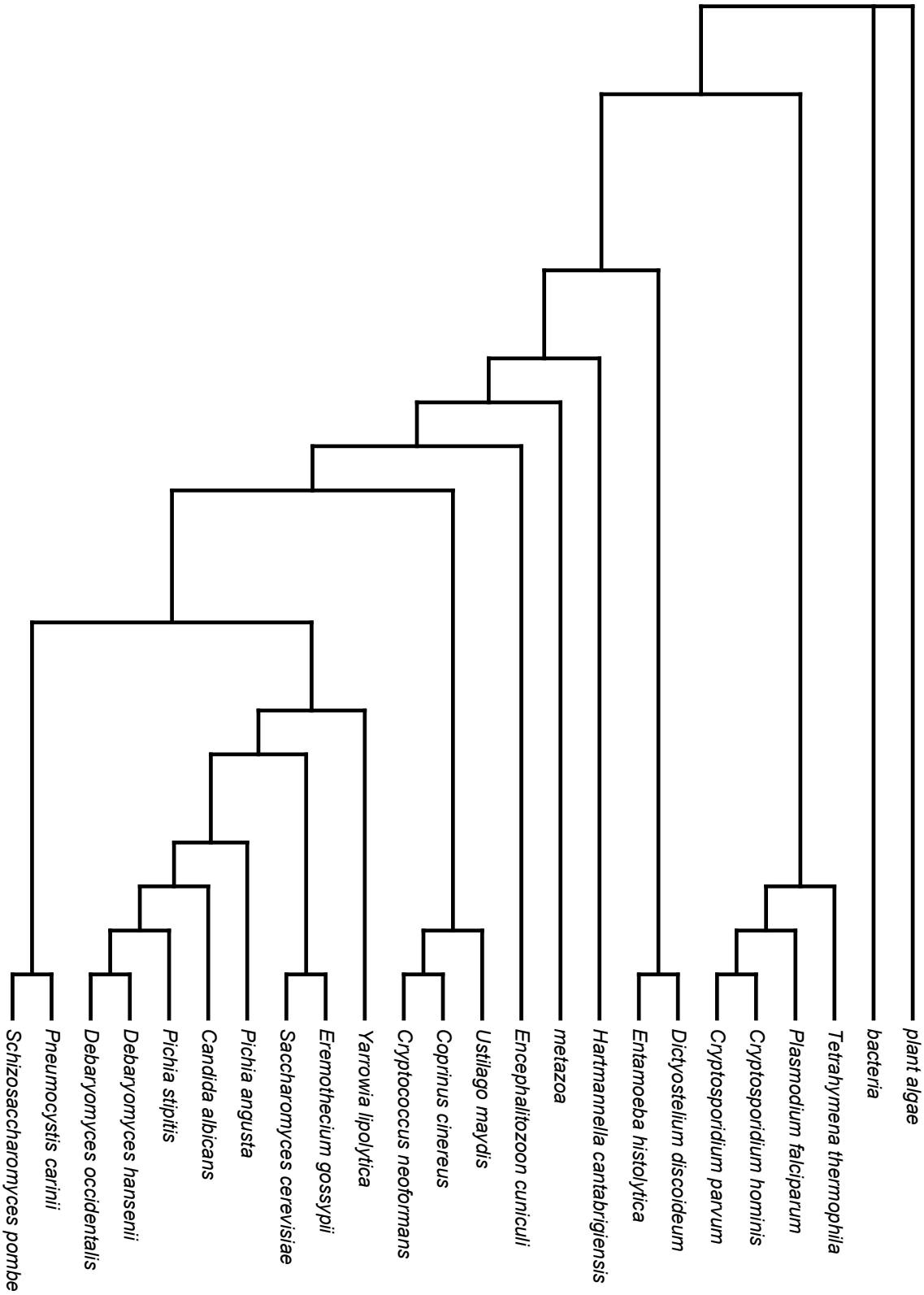


Figure 4.T.r7.s15.c.c: Round 7 subset 15 of final tree, cladogram

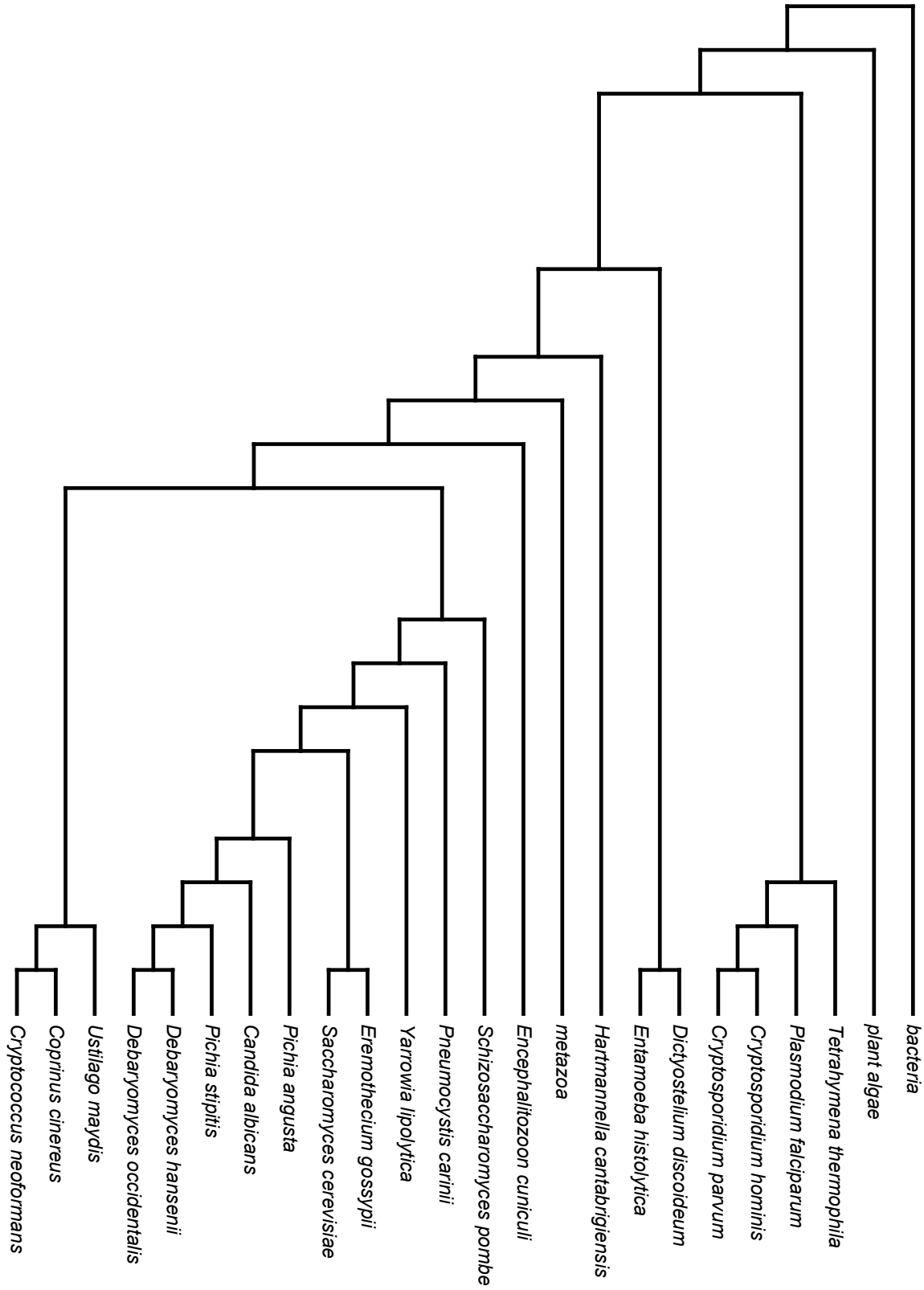


Figure 4.T.s.r7.s15.1: Round 7 subset 15, tree 1 (original) arrangement, cladogram

Final tree results

The final[481] tree results, for species with DHFR sequences only[482] except for *D. discoideum*, *E. histolytica*, and (some) outgroups, are below on pages 328-332 (in phylogram form except as noted otherwise[483]):

---

[481] By "final" is meant that:
- no rearrangements have been found that consistently improve the likelihood of the tree; and
- this was the tree used for DHFR ancestral sequence reconstruction.

[482] The tree as a whole is too large to present and be comprehensible; even the Eukaryota subset of it is likely to be too large. Given the concentration of this research, it is likely that the most reliable portions of the tree are for species with DHFR or DHFR/TS sequences known (and used); for instance, these species have been included in the most tree runs. This reliability difference is visible in the full tree in the tendency for branches involving infrequently examined species to be either very long or very short.

[483] The fungal phylogeny is displayed both with and without distances due to the short branch length of the branch to the *P. carinii*/*S. pombe* ancestral node, probably due to its association with a recent change in the phylogeny (see "Tree rearrangement for *P. carinii*, *S. pombe*", on page 320, and "Tree distances", on page 113).

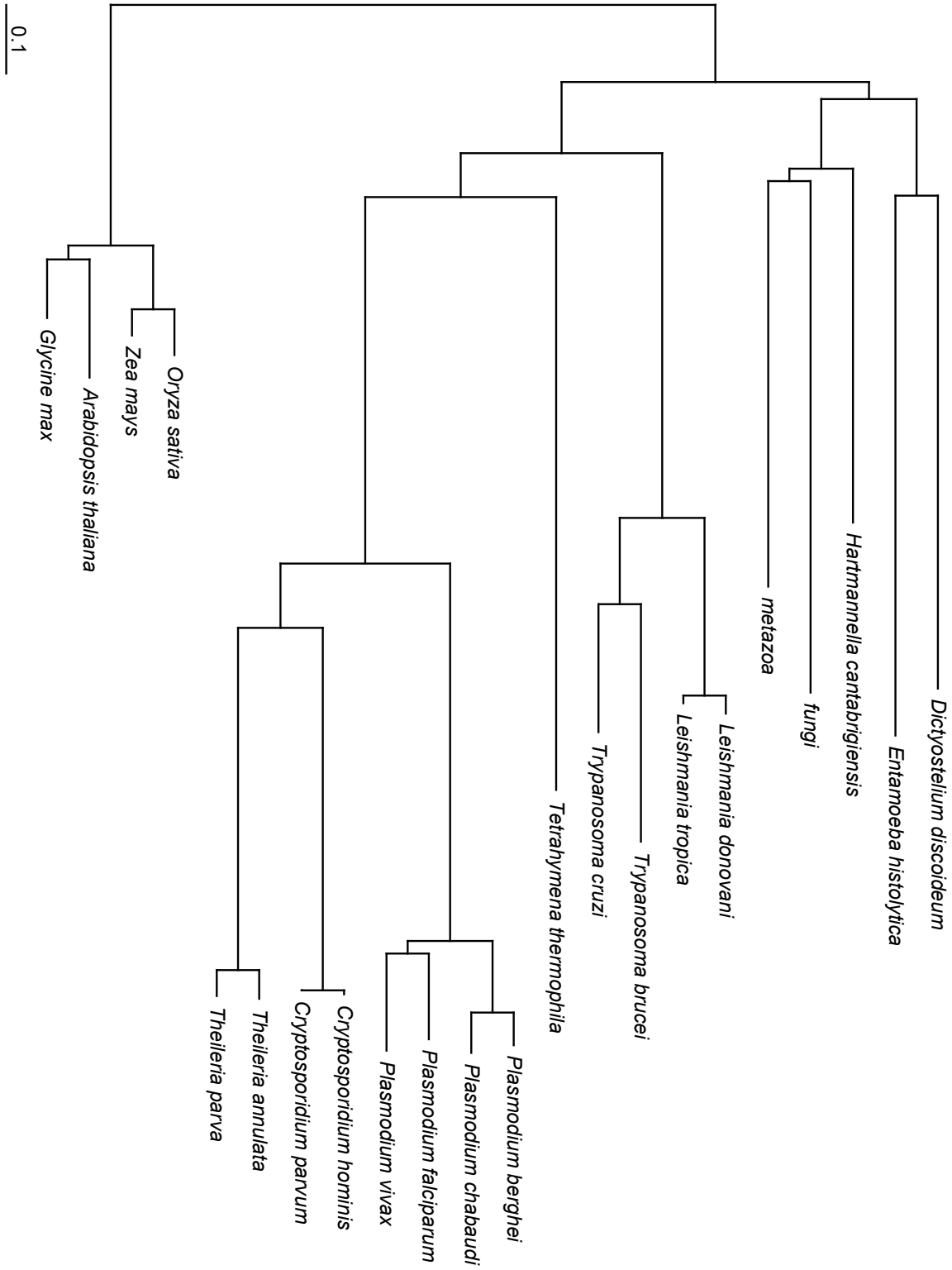


Figure 4.T.nfm: Alveolata, Kinetoplastida, Viridiplantae, and others

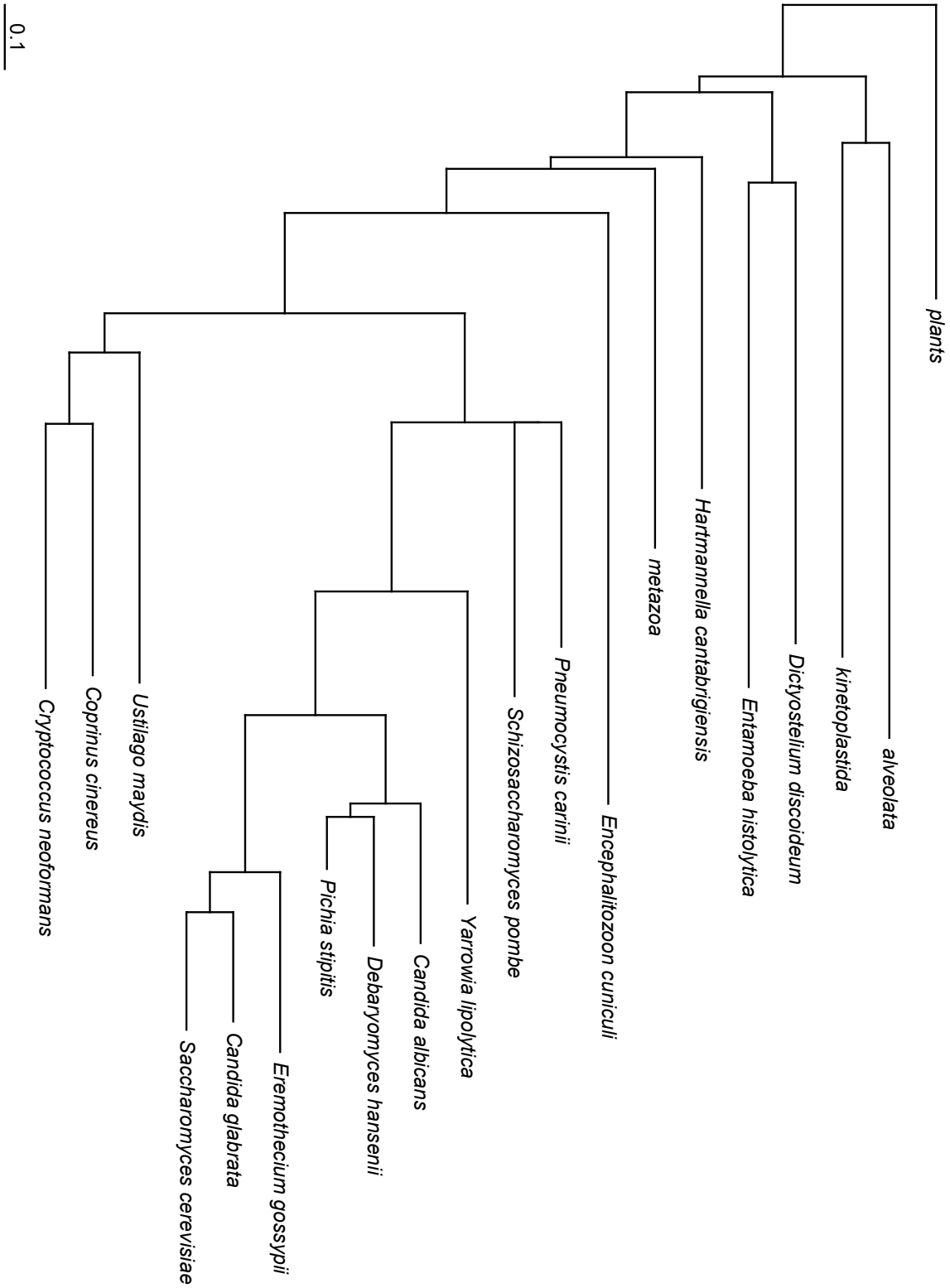


Figure 4.T.fungi.p: Fungi (phylogram)

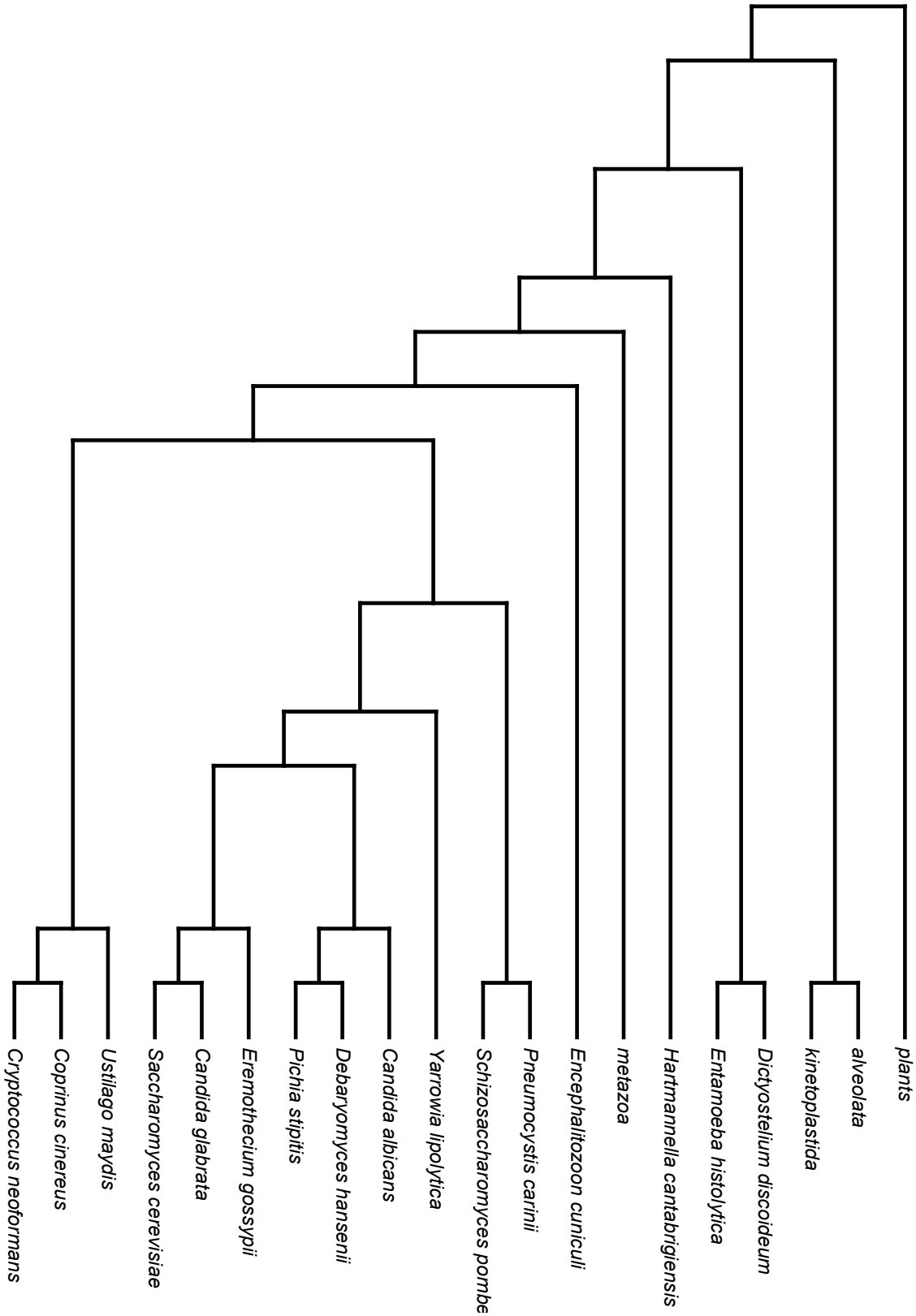


Figure 4.T.fungi.c: Fungi (cladogram)

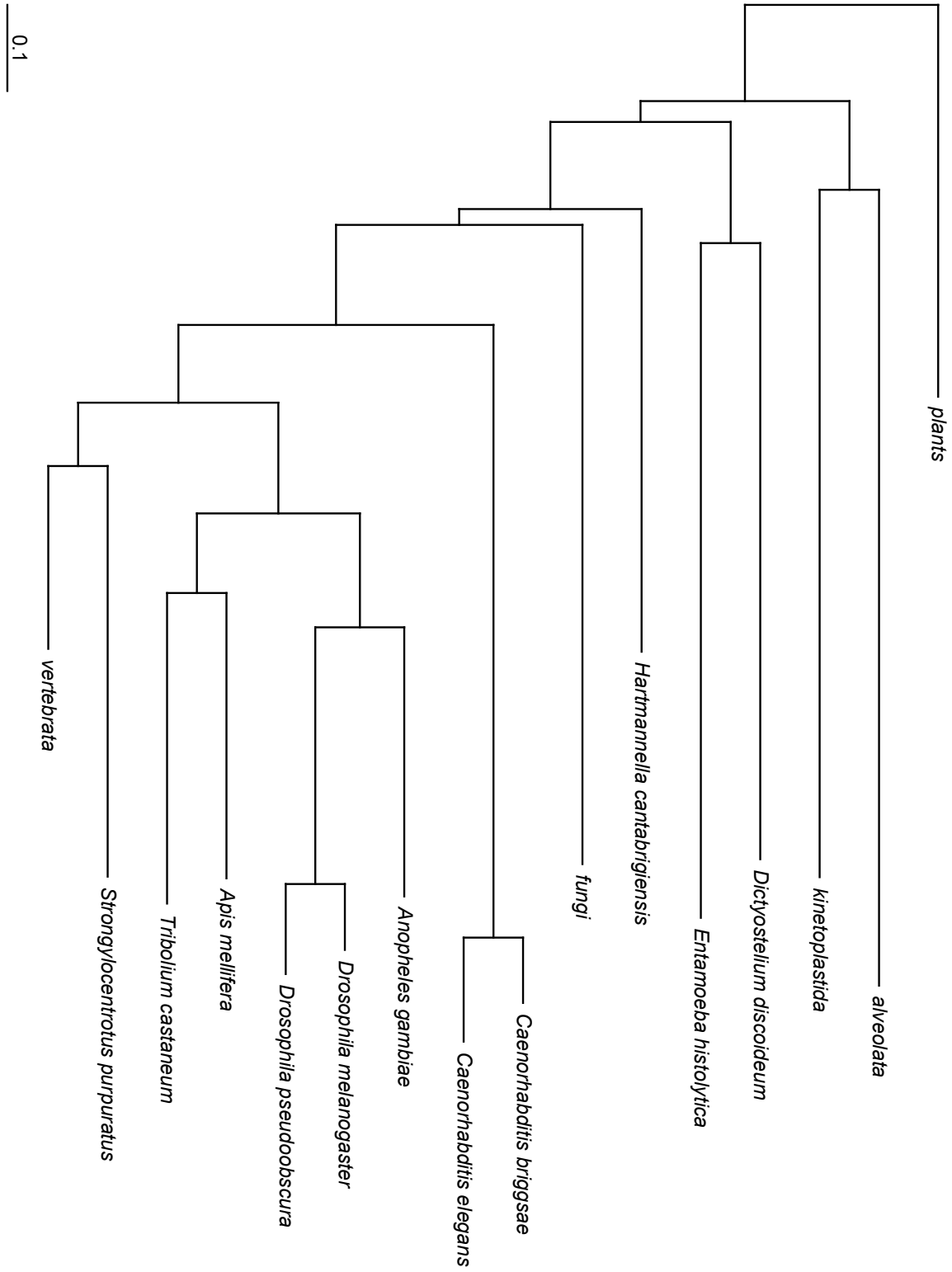


Figure 4.T.invertebrates: Invertebrates

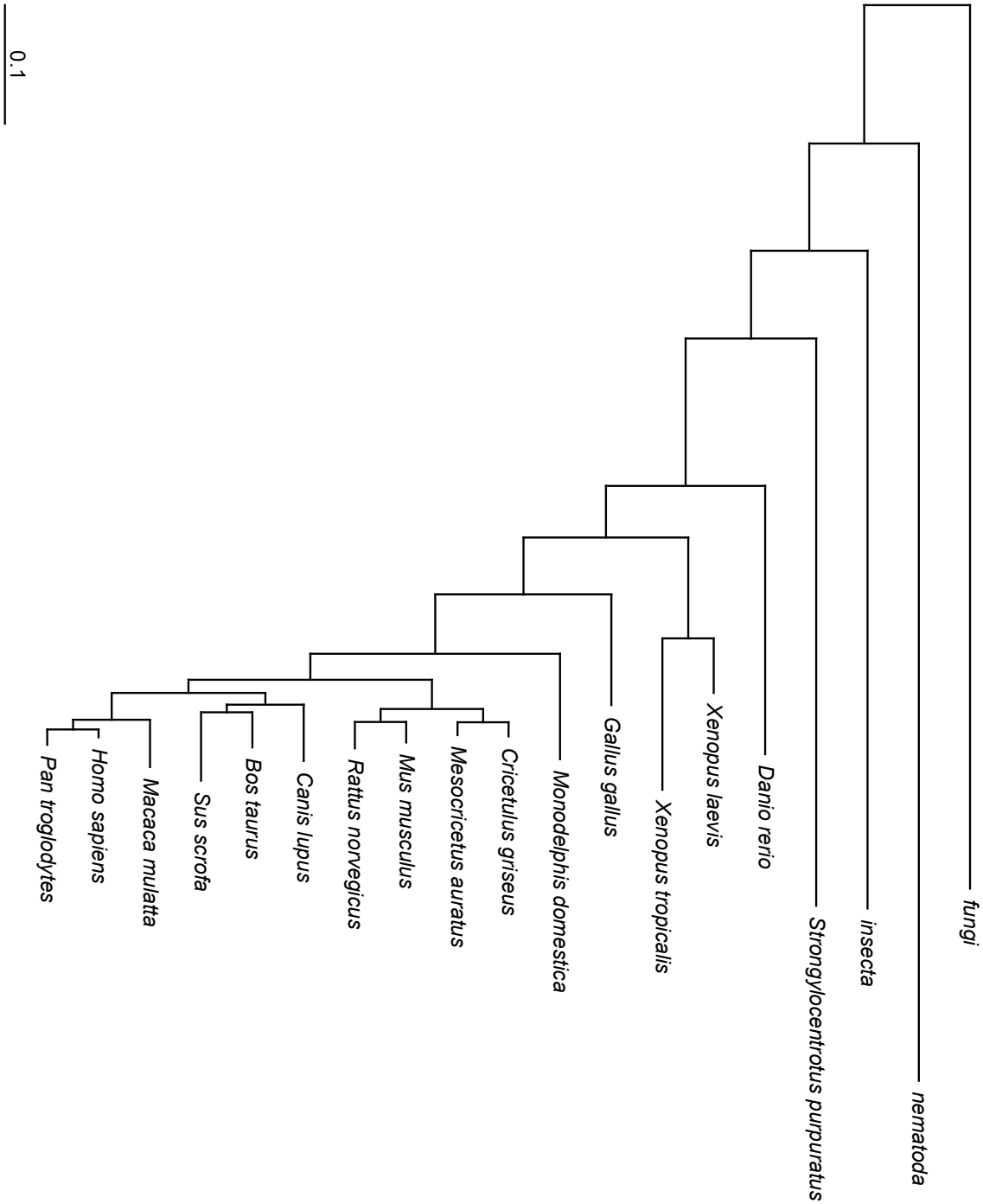


Figure 4.T.vertebrata: Vertebrata

A listing of tree files available (both as supplemental files and online) is in "Appendix L: Tree files available, cross-referenced to pictures", on page 394. All trees produced that are considered of adequate reliability will be deposited into TreeBASE (Sanderson *et al.* 1993) in NEXUS format, as will the sequence datasets used to produce them. Those considered of adequate reliability will include all trees of species with DHFR sequences used in the research, and may (depending on, for instance, the advice of committee members knowledgeable in the field) also include other trees (e.g., bacterial and/or archaeal trees). In terms of problems with the above from the viewpoint of pure phylogenetics (i.e., looking for manifestations of potential problems with respect to the models, not so much with respect to the organisms), the most obvious ones are overly long or short branch lengths. These have been mentioned previously; see “Tree distances”, on page 113 and footnote 483, on page 327, for instance.

Future work

As well as the areas noted above, one topic of interest would be to attempt to find ancestral sequences for some alternative trees, particularly for relatively recent nodes[484], then attempt the modeling of these to determine if these sequences are structurally less likely. Another area of interest is how to do more automation of the tree refinement process (also see footnote 230 under "Tree rearrangements", on page 111), for several reasons including:

- As noted under "Second round of tree rearrangements", in footnote 459, on page 252, and under "Tree search with Eukaryota (subset)", in footnote 470, on page 303, the manual editing process is prone to errors, some of which may not be caught by current programs (if they yield a technically valid tree). Better means of visualization of trees (as can be seen in the tree figures in this dissertation), and especially of visualization of the differences *between* trees (e.g., versus a "trusted" tree - see "3a. Creation of a rough starting tree", on page 192), would be of assistance. Likewise, better means of doing tasks such as rerooting, derooting, and connecting together trees - ideally in an automated[485] fashion - would be helpful.
- Currently planned (Huelsenbeck *et al.* 2007) for MrBayes is an improved implementation of constraints, in which they become probabilistic penalties on

[484] For instance, one could check to see what sequences would be derived for Urplacental and Uramniota if one were to assume that Primates and Rodentia were grouped together (closer than to, for instance, Carnivora).

[485] On the other hand, we also note earlier problems experienced with automated rerooting or derooting with MrBayes (crashes when such operations were attempted on an input user tree); this is not the least complex of problems. In some programs created locally, derooting and rerooting have been implemented, but it is difficult to do this in a reliable way that can be used in other programs.

broken groups that are supposed to be clades, instead of the current absolute bar. It is possible that this may be helpful with trying out different possible trees - but translating evolutionary hypotheses into groups is likely to pose its own usability (e.g., human error) hazards, especially when dealing with large numbers of species. Another concern will be that, as the MrBayes manual notes with regard to how to keep a tree fixed[486], constraints can be inefficient.

One thought in this regard is taking different hypotheses about the evolutionary branching order, combining them via a consensus process of some variety, and then restricting the automated rearrangements to resolving the resulting polytomies. This idea has similarities to that behind the original tree creation process in the current work (see "Initial sources", on page 72), which may perhaps serve as a warning about the complexities of consensus algorithms; on the other hand, this does have resemblances to the idea of using a "trusted" tree (see "3a. Creation of a rough starting tree", on page 192).

Some other potential improvements on the tree rearrangement/search process are related to ones that should improve the alignment of the central (not structurally aligned but outside of the 65% sequence identity) sequences. For instance, the usage of different matrices for different positions may prove helpful (Lartillot, Brinkmann, & Philippe 2007); see "Future work", on page 337, for a discussion of this in the context of alignments.

[486] Note item 5 under "MrBayes code alterations", on page 100.

5. Alignment of central sequences

The current alignment of the DHFR sequences, including the current set of postulated ancestral sequences, can be found in supplemental file "DHFR.with.fungi.2.stockholm.txt" (also available via "http://cesario.rutgers.edu/easmith/research/DHFR.with.fungi.2.stockholm.txt".)

The alignment is difficult to display, especially with additional information such as the measured solvent accessibility, due to the number of sequences and the length; with said added data, it is 302 rows (types of data) and 459 columns (note that the ".xls" format, for instance, cannot handle more than 256 rows). All portions of this alignment without an "X" in the "#=GC RF" line at the bottom of each section were done manually[487]; a brief examination of "DHFR.with.fungi.2.stockholm.txt", which is in the original format used for manual alignment, should make obvious the level of difficulty involved, making it unsurprising that there were problems with it. For a subset of some of the most important sequences, please see "Appendix K: Partial DHFR alignment", on page 384.

[487] These are considered "insert" residues by HMMER; see "Alignment using HMM", on page 129.

Future work

As well as those already discussed (e.g., see footnote 273 under “Alignment using HMM”, on page 129), several possibilities are available for improvement of the current alignment results:

- Instead of the alignment for the "insert" areas being entirely manual[488], rerun HMMER on these after having decided manually what areas are within them. This could be done either with sequences from outside the cluster present (but down-weighted considerably), or with them entirely missing. This possibility was considered during the current work, but considered too much of a headache; if the process can be more automated, then this may be a possibility.
- Further usage of the tree (Holmes & Bruno 2001), although the computational burden may be extreme.
- As well as or instead of the above, the usage of characteristics including secondary structure, intrinsically disordered nature (“nonstruct” areas), and solvent accessibility in the alignment. This could be done in several ways:
  - Modifying the likelihoods of the amino acid "mixtures" (Durbin *et al.* 1998) used in HMMer and similar programs according to the properties of the structures currently aligned to these locations (Thompson, M J & Goldstein 1996, 1997) - this is related to the similar idea of modifying the matrices

[488] In the present work, the general lesson has been learned (both with the alignment and with the tree results (see “Future work”, on page 334) that it is best to automate everything that can be automated - how manual action can improve quality is as with the alignments, by manual going over of automated results.

used depending on the properties of the positions, which may be helpful for phylogenetic work (Lartillot, Brinkmann, & Philippe 2007);

- Threading (see under "2. Phylogenetics - Ancestral Sequence Prediction", on page 7).
- Modifying the likelihood of gaps depending on the structure, as with ClustalW (Thompson, J D, Higgins, & Gibson 1994).
- Alignment using predicted secondary structures (Jennings, Edge, & Sternberg 2001; Lipke *et al.* 1995; Rice, D W & Eisenberg 1997; Simon & Simon-Lukasik 1998; Zhou & Zhou 2005) or other characteristics - see below for more.

However, these would have some difficulties:

- If done with only experimental structures, the possible changes in secondary structures, *etc.* with evolutionary time (Huang & Wang 2002; Russell & Barton 1993) may cause problems (see "Loop searches", on page 348, and "Prediction without full modeling", on page 362).
- If done with modeled structures, then while this may enable a greater degree of closeness, it may also contribute to increases in errors (on the other hand, it could also help spot errors in the models). If secondary structures are to be used[489], exactly how to define secondary structure is

[489] With regard to other predictions, it is uncertain whether a homology modeling process could accurately indicate that an area is an intrinsically disordered ("nonstruct") one. It appears likely that simulated annealing followed by energy minimization coming up with multiple different possibilities (many local minima) would indicate some such regions (ones that are only in a stable configuration when ligand-bound are among the likely exceptions). This area appears to need further research; existing work comparing intrinsically disordered areas with areas with high "temperatures" (B-factors) may be of use in this (Radivojac *et al.* 2004). If a homology modeling process indicated that an area was tightly stabilized sans interactions with other proteins (Chen *et al.* 2006; Hilser & Thompson 2007), however, this would appear to rule out a region as

another question; the present study used the classifications in the PDB files when necessary, but this is not applicable to newly created models. This question has significant uncertainty (Colloc'h *et al.* 1993; Drennan 2001), and some definitions (or a comparison of differences between the results of the definitions (Drennan 2001)) may be more informative than may other definitions.

The prediction of secondary structure (and, in some areas, of “nonstruct” - intrinsically disordered - areas) has several interesting aspects:

- One difficulty with it, as pointed out above, is defining what one is predicting (and it may be more difficult to predict some definitions as compared to others, even assuming equal usefulness of a correct prediction[490]).
- Another is that many current methods of secondary structural prediction use sequence alignments (Benner *et al.* 1997; Cuff & Barton 1999, 2000; Kloczkowski *et al.* 2002; Levin *et al.* 1993; Przybylski & Rost 2002; Salamov & Solovyev 1995; Tuckwell, Humphries, & Brass 1995), which are what one is trying to derive in this instance; there are, however, exceptions (Thompson, M J & Goldstein 1997). The same is true of the prediction of intrinsically disordered (“nonstruct”) areas (Penq *et al.* 2005), although again there are exceptions (Coeytaux & Poupon 2005).
- On the other hand, in a situation such as the present research, in which multiple structures of the same protein in widely separated species are

---

intrinsically disordered.

[490] E.g., the prediction of a "definition" that defined all structures as a "random coil" would be quite

available, this property may be usable to improve secondary structure prediction. Two possibilities in this regard are as follows:

- The usage of structurally-derived alignments with (3D) structurally known sequences and their surrounding 65% identical clusters to give further sequence alignments than are normally possible for the second possibility above (usage of alignment information for secondary structural determination). However, it is exactly those areas that we are most interested in aligning that lack good alignments even using the structural data ("uncertain" or "nonstruct" areas).
- The usage of the pattern of secondary structures - and of regions which are “nonstruct” - in common (perhaps with intervening insertions/deletions, however, as with the present alignment - see "Appendix K: Partial DHFR alignment", on page 384) between the homologous structures together with the pattern of amino acids[491]. An examination and/or prediction of properties as more continuous in nature (Andersen *et al.* 2002; Boden, Yuan, & Bailey 2006), and aligning by these (Katoh *et al.* 2002; Lipke *et al.* 1995; Simon & Simon-Lukasik 1998), may be of interest. Some examination has been made of this with regard to secondary structure with DHFR, including testing using jackknifing (leaving out proteins from a training set and seeing if predicting them is possible) with some encouraging results, but time has not permitted adequate work on this. With

---

accurate, but also quite useless.

[491] E.g., hydrophobicity; from an examination of the current DHFR alignment, it appears that hydrophobic runs are associated with DHFR's strands, which tend to be buried (Richardson, J S & Richardson 2002) - some aspects of this pattern may, however, be unusual (Schwartz & King 2006), thus enabling better prediction of DHFR's structure by keeping it in mind.

regard to intrinsic disorder ("nonstruct" areas), some prior research (Chen *et al.* 2006; Penq *et al.* 2006) indicates that having knowledge about disorder in homologous proteins may be of assistance, as may examination of intron splice sites for eukaryotes (Romero *et al.* 2006).

- An adequate prediction of secondary structure (or of other properties such as solvent accessibility) could be used to improve modeling in a number of ways[492], although there would be worries about the negative effects of incorrect predictions. On the other hand, models may help in the spotting of incorrect predictions of secondary structure (including by their not working in loop searches, *etc.*), since the full modeling process effectively takes into account much more interactions (particularly on the tertiary level) than any (other) extant means of secondary structure prediction. Whether modeling would be successful at correcting errors of prediction regarding, for instance, intrinsic disorder may be, as noted previously (see footnote 489, on page 338), another matter.
- It should be noted that it may not be necessary for secondary structure prediction to be able to indicate which of alpha-helix/beta-sheet/other a portion of structure is - it may be enough if it is able to indicate what a portion is *not*[493]. With such information, for instance, loop searches can be

[492] For instance, this could enable better loop searches by being more selective as to local secondary structure - see "Loop searches", on page 348. Note that this has already been done to some degree manually in terms of the choices of amino acid patterns used (see "Loop searches", on page 157).

[493] We strongly suspect that this would be an easier problem, partially from experience with attempting secondary structural prediction with DHFR previously and partially because of the idea of secondary structures as more continuous in nature (Andersen *et al.* 2002; Boden, Yuan, & Bailey 2006). Concerning the latter, if one is looking at either:

- probabilities or

restricted away from inappropriate sources, and alignment of incompatible areas can be avoided.

6. Determination of ancestral sequences

For the results of this, please see the alignment ("5. Alignment of central sequences", on page 336).

Gap determination thresholds

In "Gap determination", stage 2, on page 144, the threshold determination originally had been done using the state frequencies combined with what appeared to be reasonable built-in minima and maxima (e.g., 0.5 or 0.49). It was noted that the sequences for Urdeuterostomia appeared to have an unrealistically high number of gaps; structural experimentation finding that it was best (for Urdeuterostomia) to use sequences with fewer gaps (see "8. Examination of models", on page 352) was part of this conclusion. The first attempt at rewriting this section led to too few gaps, as seen in the initial fungi/metazoa[494] and Urascomycota[495] sequences. It was also realized that the exact proportion of residues assigned as gaps was actually different from the

---

- predicted degrees

of alpha helix versus beta sheet versus other, it may be adequate for some purposes to say that one of these is very low, without needing to decide between the other two.

[494] Note that the sequences used for fungi/metazoa were the 1111_* ones, with more gaps than the other set (1100_*, such as the shown 1100_SVFQ) ones tried; this was partially due to findings during attempted loop searches (see "7. Model building", on page 345) and partially due to manual examination of the alignment.

[495] These were the sequences of 10100_chars2, 10101_chars2, 10110_chars2, 10111_chars2, 11100_chars2, 11101_chars2, 11110_chars2, and 11111_chars2.

state frequencies, even with the thresholds based on the state frequencies; the reasons for this are a matter for further study.

Usage of existing models

It was generally found difficult[496] to use existing models directly[497] to deduce what sequence was more likely among several possibilities. In one instance, something similar to this deduction idea was done, however. Previously, for the Urplacental models, it was noted that residue 49 (number 137 in "Appendix K: Partial DHFR alignment", on page 384) was having problems with steric clashes, according to the MolProbity results. At the level of the Uramniota sequences, while valine was predicted as the most likely for this location, alanine was also a significant possibility (valine was estimated at 55% probability, alanine at 35%, leucine at 10%). Given the steric clashes, both valine and the smaller alanine were tried at this position, with alanine being modeled first followed by valine models derived from these. It appears that either is possible, although alanine may be somewhat more likely from the modeling results.

Another respect in which existing models were used was in looking for residue correlations that were correlated with structural closeness. As well as

[496] One reason for this difficulty was the number of changes taking place at once, particularly with regard to gaps and other changes potentially significantly affecting the backbone configuration. Doing more modeling stages with fewer changes per stage (see "7. Model building", on page 146) may be of assistance with this, especially if it can be automated to a greater degree (Aszodi, Munro, & Taylor 1997; Azarya-Sprinzak *et al.* 1997; Blundell 1991; Bowie, Luthy, & Eisenberg 1991; Fornasari, Parisi, & Echave 2002; Koehl & Levitt 2002; Ponder & Richards 1987a, 1987b; Sunyaev *et al.* 1997; Wilmanns & Eisenberg 1995; Word *et al.* 2000). Please see "Prediction without full modeling", on page 362.

experimentally determined structures, existing models were incorporated into this check (see "Sequence determination", on page 135).

Discussion and future work

It appears that the sequence determination portion of this went reasonably well (although the usage of residue correlations needs to be further automated); the gap determination portion needs further work, both in terms of automation and in terms of determining, for instance, what coding of gaps is the most valuable. The difficulties with the DHFR alignment, particularly in those areas not structurally aligned, may have contributed considerably to this problem. It is suggested that an alignment containing less non-"struct" areas (e.g., ADH1 or myoglobin - the latter being a protein with some evolutionary correlations already known (Dutheil & Galtier 2007; Neher 1994)) would be valuable for usage, along with modeling, as a way to test different gap inference procedures on a more reliable alignment. Alternatively, studies such as that used for TipDate (Rambaut 2000) of recent evolution on a known tree and known ancestors may be preferable, if reasonable alignments can be derived - as appears likely to be the case - although whether such a study's sequences will include sufficient gaps for checking gap determination is questionable. This study's sequences are, however, likely to be of use in, for instance, checking the effects of the alterations to MrBayes for handling polymorphism (see item 2 under "MrBayes code alterations", on page 98).

[497] By "directly" is meant without going through at least some level of modeling (or attempt at modeling).

As well as the problems with the DHFR alignment, an additional problem was in properly translating gap positions, since the gap prediction results are *in respect to* a set of positions (frequently non-continuous with respect to the alignment as a whole), that were in the gap partitions[498] used for ancestral "sequence" prediction. This difficulty was especially a problem when attempting to combine the automated determination of correlations and manual examination of the alignment, as sometimes had to be done with the sequence determination in insertion areas (e.g., the loop area mentioned on page 173) due to limits on the number of sequences practically modelable.

7. Model building

In terms of results, one result of this work has been the production of a (only partially automated, and not very reliable as yet) set of homology modeling programs. Perhaps most importantly, unlike all other such program sets of which we are aware, all of the locally produced programs are open-source and draw chiefly on programs that are themselves open-source (e.g., GROMACS). Admittedly, the set of programs in question is only partially automated, and cannot be said to be very reliable yet, but it may be the basis for further work in this area. Also with regard to future work, as well as that below and other material, please note under "Simulated annealing when needed", item 4, on page 186.

Concerning models and sequences, it should be noted that *multiple* models have generally been derived for each sequence. While most commonly these are from different levels of minimization with different degrees of restraints (see "Creation of restraints", on page 170), in some cases models have been derived using different:

- sources (e.g., group1 versus group2 for Uramniota - see "Appendix E: MolProbity results", on page 371); or
- techniques (e.g., restraining to a rotamer library or searching through all possibilities for a starting rotamer position - again, see "Appendix E: MolProbity results", on page 371).

One respect in which multiple models have been used is to have multiple templates on which to base the next level of modeling, then using whatever portion of the templates appears to be the most valid and/or averaging them together - see "Assignment of initial coordinates", on page 150.

In the course of model building, it was determined that some predicted sequences did not appear to be physically realistic, at least in the context of the existing structure (even with flexibility via minimization). At the fungi/metazoa common ancestor stage, the 1111_chars2[499] sequence, when the second stage (conjugate gradient minimization) of partially frozen minimization was reached, exhibited a behavior known as an "exploding simulation" (Lindahl *et al.* 2007). In this, a warning was emitted of two atoms that were supposed to be close by

[498] These would be either binary or “DNA” - see “Gap determination”, on page 139.
[499] Note that this indicates that the "1111" gap arrangement was used with the most likely sequence according to MrBayes ("chars2"; see "Sequence determination", on page 135).

(being within 4 bonds of each other) being beyond the maximum range for such. In this instance, the atoms in question (the backbone nitrogens of valine 160 and aspartic acid 161)[500] were at a distance of 1,281,220,157.935 nm; given that this distance is many orders of magnitude greater than the diameter of the entire molecule[501], this situation does not qualify as physically realistic. The most likely reason for this error appears to be the neighboring residue (162, or 329 in "Appendix K: Partial DHFR alignment", on page 384), which is a tryptophan in 1111_chars2 (as predicted as most likely by MrBayes) but a tyrosine[502] in the other sequences. One reason for believing this, as well as the bulky nature of tryptophan (the largest amino acid), is that the 1111_STF sequence gave some problems[503] with the tyrosine at 162 (329), although these were hopefully solved in later minimization work. The full kinemages (Richardson, D C & Richardson 1992) output by MolProbity on attempted analysis were not visualizable in KiNG (Richardson, D C 2007) due to the large number of errors found, and the "thumbnail" graphical capabilities were not reliable (causing intermittent browser problems). The PDB-format file for this structure is fungi_metazoa.1111_chars2.idm.freeze1.new.reduce3.ent (see "Appendix M: Model PDB-format files", on page 403).

---

[500] In the alignment in "Appendix K: Partial DHFR alignment", on page 384, these are at 322 and 324, respectively.

[501] 1,281,220,157.935 nm is over 1 meter.

[502] The probabilities given by MrBayes were 53% for tryptophan, 29% for tyrosine, and 18% for phenylalanine. Tyrosine was decided on manually based on that 154 (319 in "Appendix K: Partial DHFR alignment", on page 384) was a lysine (probability over 0.95 - treated as a probability of 1) and no cases whatsoever were seen of a lysine at this position and a tryptophan at 162 (329).

[503] To be precise, the same error type showed up as with the tryptophan, but with the atoms involved (a ring carbon (CD2) and the ring OH oxygen) being too far apart (1.014 nm) at one point during the minimization. After rerunning with slightly higher tolerances (a "table-extension" of 1.2 nm), this problem was solved (no further such errors were seen, even in later stages sans modifications of tolerances).

Loop searches

In the loop searches done as a part of this research, instead of searches for geometrically-conformant (to the existing structure) loops being done, ones for loops conformant to the desired *sequence* have been done. One advantage of this methodology that has been found is that, if a particular sequence cannot be found within a (known) 3D structure, then this may indicate a lack of realism in the postulated sequence (due to a *possible* inability of proteins with (known) structures[504] to accommodate it). In the present study, this happened with the fungi/metazoa sequence for 1100_chars2 and 1100_SVYQ (and other sequences with the "Y" component)[505].

Another instance of this, not recognized at first as such, was with the Urdeuterostomia 0010_* sequences. While a loop search apparently was successful[506] initially, attempted vacuum energy minimizations (see "Non-frozen

[504] Admittedly, this leaves open the possibility that the sequence may not have been found due to one or more of:

1. Overly strict constraints on the residues accepted - this was partially allowed for by examining scanprosite's output as to the expected number of matches (Nicodeme 2001) and making sure it was at least 1 (normally, even for cases where a sequence was not found, it was at least 10);
2. A chance lack of studies on sequences matching the criteria (although, given that the sequences of interest are in a protein resembling one of considerable research interest as indicated by multiple structures, this seems dubious); or
3. A tendency for the area in question to lack structure (under many conditions) and thus not be seen - in other words, this would suggest that this might be a "nonstruct" (intrinsically disordered) area in the protein. Such a possibility might be examinable by looking at the properties of such areas further (Alroy 1995; Chen *et al.* 2006; Coeytaux & Poupon 2005; Penq *et al.* 2005; Penq *et al.* 2006; Romero *et al.* 2006).

[505] This was also part of what led to the further examination of the postulated longer loop insertion present in the 1100_* sequences for fungi/metazoa (see positions 332-353 in "Appendix K: Partial DHFR alignment", on page 384; this is the area for which a loop search was unsuccessful). This loop was concluded to be a likely later insertion in the fungi.

[506] Indeed, the loop search, as far as can be told currently (see "8. Examination of models", on

vacuum minimization", on page 174) of these structures using the conjugate gradients minimizer (following the initial minimization using the steepest descents minimizer) were unsuccessful, yielding:

1. Either no change at all[507] or very few rounds of changes (e.g., 2).
2. A final maximum force that was either:
    a. well above the desired threshold (e.g., $6.702*10^{15}$ kJ/(mol*nm), a value difficult to achieve by chemical processes); or
    b. Obviously invalid (a force of 0 and a potential energy of "nan" (see footnote 507, on page 349)) for computational reasons.

Initially, this was thought to be due to clashes between aspartic acid 228 (by the alignment in "Appendix K: Partial DHFR alignment", on page 384) and neighboring residues[508]. However, a further examination of the results (via MolProbity, including its "thumbnail" visualization capabilities) indicated that the actual problem was with the regions with loop searches, as indicated by considerable gaps between portions of the protein[509]. As with the above (on page

---

page 352), appears to have been successful for the corresponding 0011_* sequences.

[507] For 0010_KD, the minimization was unable to do any changes, and printed out a coordinate file with "nan" (Not a Number - a computer error similar to a report of "infinity", but not implying great - or, for "negative infinity", small - size) in place of numbers. The available structural file is thus that from the first round of vacuum minimization, with the steepest descents minimizer.

[508] If so, then the alternative glutamic acid would have been preferable due to being longer and thus more conformationally flexible (and able to accommodate other residues with branching at a lower level, such as the neighboring leucine).

[509] As well as the sequence itself being not physically realistic (note regarding Urdeuterostomia under “Gap determination thresholds”, on page 342), another possibility is that the difference in alignment at 234-235 (see "Appendix K: Partial DHFR alignment", on page 384) and resultant difference in position for the loop search for said residue resulted in a problematic loop insertion. Such a problematic insertion could pull the other areas such that the energy minimization could not reconnect the ends, as it was able to do with the 0011_* sequences. However, given that the alignment area in question is the *reason* for the identity of the sequence at 234-235 (leucine or aspartic acid, depending on to what other residues it is aligned), this can also be classified as a problem with the sequence.

347), the kinemages are not usable and other forms of visualization seem likely to not show any more useful information. The PDB files of these structures are:

- ascomycota.0010_KD.vacuum.new.reduce3.ent;
- ascomycota.0010_KP.vacuum2.new.reduce3.ent;
- ascomycota.0010_QD.vacuum2.new.reduce3.ent; and
- ascomycota.0010_QP.vacuum2.new.reduce3.ent.

Please see “Appendix M: Model PDB-format files”, on page 403.

For future work on loop searches, a much higher degree of automation is desirable[510]. A combination of the current loop search method[511] with ones looking more at the geometry of the anchor elements - or, perhaps better, at the solvent accessibility and other environmental characteristics (Topham *et al.* 1993; Wohlfahrt, Hangoc, & Schomburg 2002), to better take into account backbone flexibility - may also be indicated. Another possibility is (for the above and other purposes, such as secondary structure prediction) to examine the likely solvent accessibility of the *area*, not necessarily of the residues themselves, since a residue loop can "flip" during the course of evolution. Such a “flip” happened in the present study, for instance, for the fungi/metazoa models at

[510] Please see the Dedication, on page vi, for someone who needs to be thanked (again) for doing (painfully boring) manual loop searches (Engel 2007) to assist the author.

[511] Another possibility for searching would be using a “profile”-type search - as per, for instance PSI-BLAST (Altschul *et al.* 1997; Schaffer *et al.* 2001) - using:

- the ancestral sequence probabilities;
- a matrix (e.g., the “Nussinov” one) to allow for some possibility of different amino acids, e.g. as per the option with HMMER to use a PAM matrix (Durbin *et al.* 1998; Eddy & Birney 2003); or
- a combination of these (most needed with positions where it was desired to try more than one possible ancestral sequence, so that one could eliminate the alternatives that would be being tried in alternative sequences).

Examination of some research going in the reverse direction (Fornasari, Parisi, & Echave 2002; Koehl & Levitt 2002) may be of interest.

residue[512] 86 by the Amniota (with 3D structures known) numbering[513], with a formerly buried alanine mutating to a fully exposed aspartic acid, while the next residue became buried; this is shown by the solvent accessibility information in the alignment (see "5. Alignment of central sequences", on page 336).

Rotamer searches

One suggestion for future work concerning rotamer searches is checking, in an automated manner, on the beta carbon rotation of any existing residue and using it as the basis for, at the minimum, deciding on which set of rotamers for a mutated residue to try. This idea would be especially applicable for cases without differences in branching at the beta carbon.

The possibility of applying rotamer searches to bad rotamers or beta carbon deviations detected by MolProbity (see "MolProbity", on page 186) should be explored, although we hope that some improvements in the determination of the initial positions will help with this. Under exploration is the idea of taking model residues with idealized beta carbon positions (Lovell *et al.* 2003) and aligning them in for each residue as, at minimum, one of the contributing templates[514] for the beta carbon.

---

[512] This area was loop modeled using the *Plasmodium* and *Cryptosporidium* DHFRs, which also have a charged residue at this position; the alteration in question is accompanied by changes in prolines and gaps surrounding it.

[513] This is position 191 in the alignment in "Appendix K: Partial DHFR alignment", on page 384.

[514] Alternatively, using the location thus determined as a replacement for the beta carbon may be preferable, especially if the other templates appeared to have beta carbon deviation (as detected by MolProbity - see “MolProbity” on page 186) or similar problems.

Rotamer searches were also used for some models (at the Uramniota stage) to attempt to relieve steric clashes, at tyrosine 177 (with 136-138)[515] and serine 42. It was, however, concluded that energy minimization together with simulated annealing (with better restraints; see under "Creation of restraints", on page 173) was a more efficient way to fix such problems.

8. Examination of models

The MolProbity results are online at http://cesario.rutgers.edu/easmith/research/molprobity/; please see below (on page 355) and in "Appendix E: MolProbity results", on page 371, for a summary. Information on model files in PDB format can be found in "Appendix M: Model PDB-format files", on page 403.

Concerning Uramniota, both the MolProbity examination of the results of attempted modeling (see "Appendix E: MolProbity results", on page 371) and the results of modeling indicated that residue 275 (in "Appendix K: Partial DHFR alignment", on page 384; this is residue 129 in the Uramniota model) was a glycine, not an isoleucine. If this residue was an isoleucine, this resulted in increased clashes between residues directly after it (283-285, or 136-138 in the Uramniota model) with residue 387 (177 in the Uramniota model). The sequence distance between these residues - 387 is near the end - points out how much the effects of a glycine/non-glycine difference can be non-local, at least in terms of

[515] In "Appendix K: Partial DHFR alignment", on page 384, 136-138 are positions 283-285; 177 is position 387.

sequence. Notably, the glycine at 275 (129 in the Uramniota models) in these models has phi/psi angles that are extremely uncommon for non-glycines (Lovell *et al.* 2003):

Sequence identifier	Model[516]	Phi deg.	Psi deg.
AGA	full	122.17	163.27
	full2	123.31	163.58
PGV	full	115.31	-146.9
	full2	115.15	-144.36

Exactly how the chicken structure functions with an isoleucine at this position[517] is a question for further investigation. It is presumably through alterations in other residues; the most likely candidate appears to be the glutamine (in the chicken sequence) immediately after 387, which is a glutamic acid in the Uramniota models (as with most Deuterostomia) - a lack of charge repulsion appears likely to play a role in the chicken structure.

With regard to Urdeuterostomia, some choices with regard to what gap and sequence arrangements to use were from MolProbity results (see the supplemental files referenced in "Appendix E: MolProbity results", on page 371). In this, the first few sequences (e.g., "K_D" and "K_K", shown in the sequence alignment in "Appendix K: Partial DHFR alignment", on page 384) were modeled, followed by both comparisons between their results and comparisons with the results of adding new sequences (based on these models and on the templates from the earlier level). From an examination of the MolProbity results, it is concluded that the practice of basing models on models at the same "level" was

[516] The “full” and “full2” versions differ on the minimization conditions; see “Appendix E: MolProbity results”, on page 371, for more information.

probably not a good idea, due to the accumulation of errors, despite the (considerable!) timesavings in rotamer searches, although it is still estimated that these models may be of some use[518]. Further filtration to use only, for instance, the rotamer results from existing models that appear problematic (as opposed to all positional (xyz coordinate) information) appears to be recommended. Also examined at the time, and a large part of why the above procedure was continued despite the MolProbity results, were the number and degree of restraint violations, using GROMACS' "g_disre". That these results (in particular those for intra-DHFR restraints, not those for DHFR-NADPH) appeared to be improving - despite the MolProbity results' negative trend - is a large part of why the validity of the current restraint system appears dubious (see "Creation of restraints" on page 170).

Unfortunately, the quality of the models, as measured by MolProbity (see "MolProbity", on page 186), progressively declined - to an unacceptable level at the fungi/metazoa common ancestor stage and afterward (and the Urdeuterostomia stage may be considered debatable). A summary is on page 355; see "Appendix E: MolProbity results", on page 371, for more details. This problem was probably due to a combination of:

1. ancestral gap prediction problems (including due to alignment problems) - see above; and

[517] A similar question appears likely to be present for other Aves.
[518] As well as the current "final" sequences used, the "EEK" and "_EEK" models appear possible from the MolProbity results. It is suggested to try minimization without restraints other than on the NADPH on all of these. On the other hand, current efforts at applying simulated annealing to the currently predicted fungi/metazoa ancestral sequences appear promising; a summary is on page

2. cumulative errors in modeling[519] that were not corrected due to time pressure.

Stage[520]	Used further?[521]	Bad Rotamer %age Range	Bad Phi/Psi %age Range	Good Phi/Psi %age Range
Urplacental	No	2.96-3.55	1.63-2.17	87.5-89.67
	Yes	1.78-4.73	0.54-4.35	88.04-91.85
Uramniota	Yes	1.18-3.57	0.54-1.63	87.5-95.11
Urdeuterostomia	Yes	4.27-4.91	3.37-5.08	84.75-88.2
Fungi/Metazoa, Try 1	No[522]	1.78-2.94	3.23-4.92	76.5-78.14
	Yes	2.96-5.29	3.28-4.37	78.69-83.61
Fungi/Metazoa, Try 2	Yes (Planned)	1.78-5.33	2.19-3.28	82.51-89.62
Urascomycota, Try 1	No	6.99-9.14	5.85-7.32	76.1-78.05

The MolProbity result ranges[523] for experimentally determined structures (original and, when applicable, minimized), are summarized below for comparison:

Species	Minimized?	Bad Rotamer %age Range	Bad Phi/Psi %age Range	Good Phi/Psi %age Range
Homo sapiens	No	14.29-18.79	0-1.11	94.44-96.74
	Yes	2.38-2.98	1.09-1.09	91.85-92.39
Mus musculus[524]	No	9.52	1.09	92.93
	Yes	4.17-4.76	3.26-3.8	91.85-91.85
G. gallus	No	5.42-5.56	0-0	96.2-96.74
	Yes	0.6-1.81	0.54-0.54	89.13-89.67

355; see "Appendix E: MolProbity results", on page 371, for more details

[519] This includes errors due to:
1. a need for program improvements (e.g., with regard to restraints);
2. the failure to compensate/correct (such as through simulated annealing with *appropriate* restraints) for previous errors;

[520] Please see Figure 3.4, on page 149, for information on the phylogenetic location of each stage.

[521] "Used further?" is for whether the models summarized were used as templates for most residues in the next stage; note that only sequences for which at least some models were used are summarized.

[522] These were the results from the first attempt at simulated annealing; they were not adequately alignable (see under "Simulated annealing when needed", on page 184) to each other or to the non-annealed sequences. This problem was probably partially due to an overly high temperature (see "Simulated annealing when needed", on page 183) and partially due to bad restraints (see "Creation of restraints", on page 170). (Remaining to be checked is whether they are alignable to any of the new simulated annealing results ("Fungi/Metazoa, Try 2"); .this appears unlikely, however.)

[523] If only one value is given, only one structure of this type was checked; in most cases, no such structure exists.

[524] Note that there is only one mouse DHFR structure (1U70); it is a mutant, and the quality of the structure (in terms of the analysis results) may have been reduced by this.

Species	Minimized?	Bad Rotamer %age Range	Bad Phi/Psi %age Range	Good Phi/Psi %age Range
P. falcip.	No	2.83-4.65	0.45-0.91	94.14-96.77

Future work

As noted on page 355 as "Fungi/Metazoa, Try 2", an attempt has been made to improve the existing Fungi/Metazoa models ("Try 1"), via simulated annealing (using non-strict restraints, followed by energy minimization with restraints only on the NADPH). These new models may be usable as the basis for a new Urascomycota set of models, especially with better programs for the stage of coordinate assignment (see "Assignment of initial coordinates", on page 150, and "7. Model building", on page 345). On the other hand, it may be more valuable to investigate better restraints (Flohil, Vriend, & Berendsen 2002) and vacuum force fields (Summa & Levitt 2007) first. It may also be preferable to do a better version of the Urdeuterostomia models (including the addition of other sequence possibilities, to be evaluated via modeling).

Final evaluation

When one or more *Pneumocystis carinii* models are determined, their quality will be examined via automated structural alignment under blinded conditions (no human examination, except for the RMSD and number of residues aligned) to the actual *P. carinii* structures, and compared with the results of alignments between the different *P. carinii* structures. No further examination will take place at this time, since the *P. carinii* models are not the intended end goal, and further examination of a fungal structure would potentially bias the subsequent attempt

to model the *C. albicans* structure; the results will be used solely for feedback on how successful the modeling was. Later, once the *C. albicans* models are created, a full examination of the models and comparison of them with the existing structures will take place.

Summary of progress

1. The most well developed aspects of this research project are the two databases that have been developed:
    a. One of manually reviewed structural alignments[525], plus sequence alignments for sequences adequately close for sequence alignment to be trustworthy. (See "3b. Alignment of other sequences", on page 78.) This database includes information on areas that are not structurally alignable[526], due to being either:
        (i) missing in the structural files (intrinsically disordered areas (Le Gall *et al.* 2007)); or

[525] Study of the properties of the structural alignments themselves, such as for any phi/psi correlations, may be of interest.

[526] Moreover, the patterns of amino acids, etc., for these "nonstruct" and "uncertain" areas may be of interest in and of themselves - see footnote 177 under "Evaluation of structural alignment reliability", on page 86. Some prior research has been done regarding each of these in terms of amino acid composition (Chang, M S S & Benner 2004; Coeytaux & Poupon 2005; Penq *et al.* 2005; Penq *et al.* 2006); the question of whether such areas display differing evolutionary patterns in other respects (e.g., whether one should use different matrices with them) is an open question. (Admittedly, it would be difficult, by definition, to study these over long ranges - at least for areas of significant size - since this would necessitate structural alignment. However, it may be possible to study the evolution of ligand-binding intrinsically disordered areas (Chen *et al.* 2006; Hilser & Thompson 2007) by studying their structures in the bound state, although their sequences may not diverge sufficiently in the ligand-binding area unless the ligand had also altered.) Note that any deductions about the "uncertain" areas that may be made would also potentially help in automating the procedure for deciding what areas are "uncertain" - this would be particularly of value for portions of the database purely using the HOMSTRAD alignments, which do not contain this information, and for further expansion of the database. (It may be of interest to annotate the database with information about the origins of the structural alignments.) Prior research into determining areas of uncertainty in *sequence* alignments may also be of

(ii) uncertain in the alignment even with structure.

This database focuses on proteins that are of functional interest[527] and/or are found in organisms of evolutionary interest (and thus should be of use in further evolutionary studies). This database is likely to be of long-term use, both in future work in the area of this research (as mentioned elsewhere, e.g., under "Discussion and future work", on page 344) and in other work.

b. One of structures (chains in PDB files) versus species names, with manual review when other sources conflict[528]. (See "Database of structures and species", on page 55.) The primary usefulness of this database is likely to be in notifying other databases of errors; it is intended that an automated mechanism to check other databases for incongruities relative to it will be created. (A summary of these findings may be of interest as an advisory[529] paper.)

2. Enhancements have been made on the open-source LSQRMS program (Alexandrov & Graham 2003) for doing structural alignments (Gerstein & Levitt 1996, 1998), including the usage of all heavy main-chain atoms instead of simply the alpha carbon (or beta carbon, in some variants used by the Structal method's authors). Open-source "wrapper" Perl programs

---

interest.

[527] The database may thus be of use in studying the levels of selective pressure and variability within these proteins, the degree of correlation of mutations, and potentially what mutations (to, for instance, induce thermophilicity) may be best for these proteins.

[528] I.e., a database that resolves nomenclatural issues for PDB files. Note that these issues go beyond the level of genera, or even somewhat larger levels - see the example of 1KLK on page 55, for instance.

[529] The paper's intended audiences would include both PDB depositors/curators and other users of structural data.

have been constructed that aid in the evaluation and usage of the modified LSQRMS program. Further improvements may be made on this, with comparisons versus the structural alignment database, by - for instance - the implementation of gap penalties[530] for the structural alignment.

3. A methodology of creating an initial starting tree from one (or possibly more) "trusted" tree(s) (see "3a. Creation of a rough starting tree", on page 192) and a collection of other trees, weighted for accuracy according to the "trusted" tree(s), has been explored. Phylogenetic supertrees are an area of general interest (Bininda-Emonds, Gittleman, & Purvis 1999; Bininda-Emonds, Gittleman, & Steel 2002; Fitzpatrick *et al.* 2006; Moret *et al.* 2003; Piaggio-Talice, Burleigh, & Eulenstein 2004); the weighting method used appears to be new, and is suited to automated usage with a database such as TreeBASE (Sanderson *et al.* 1993). This method may also be usable to extract possible tree rearrangements on a more automated basis (see "Future work", on page 334), by looking:
    a. for areas in which studies conflict; and, eventually,
    b. for areas of polytomy in the "trusted" tree(s) for which no studies have been deposited in the database.
4. Some minor enhancements have been made on the (open-source) program HMMer that may be of use to others. This may especially be true

[530] Note that the Structal method's original authors have explored this area to some degree. The existing explorations use secondary structural data, however, and this creates a dependence on the exact definition of secondary structures in use (Colloc'h *et al.* 1993; Drennan 2001).

since it has been done along with the implementation of a tree-weighting scheme allowing the external input of a tree with distances.

5. Some enhancements have been made on the (open-source) program MrBayes. These, with further verification work in some cases to examine how advantageous they are (and in what circumstances they are advantageous, if this varies), should be of use for future work in this area and to other users of MrBayes.
6. Some potentially interesting phylogenetic findings have been made, although most[531] are in need of extensive review and further study. If it is concluded that some of these are in error for reasons of interest[532], then, because the dataset used contains structural information, the structural causes of said errors can be examined. For instance, if the errors are due to correlated mutations, then the structural locations of the mutated residues can be examined. If the errors are due to parallel or convergent evolution, then the likely functional correlates of the mutations will be easier to deduce with structures available.
7. A start has been made on the creation of an open-source suite of modeling programs. These programs already include capabilities for using multiple templates and for handling multiple models in parallel. The idea of avoiding rotamer searches by taking into account existing sidechain orientations (even if the sidechain is not the final desired one) and/or the

[531] It is suggested that, particularly given prior evidence (Fitzpatrick *et al.* 2006; Sugita & Nakase 1999a, 1999b), the placement of "*Candida" glabrata* in the *Candida* genus is a taxonomic error requiring rectification.
[532] I.e., not due to human error.

results of loop searches appears to be a new one, and may (with an improved implementation) be of wider interest.

8. Further examination of the existing models and problems with some of them (including the causes of deciding against some sequences being realistic ancestral sequence possibilities) may lead to conclusions further of interest. Cases in which models "blew up" due to sequence problems may be of interest for future work in finding ways for structural data to help in phylogenetics; see "Prediction without full modeling", on page 362.
9. The models down to at least Uramniota, and perhaps Urdeuterostomia, appear to be good (although further examination is desirable). Further work with these may enable to completion of the original aim of the project. Moreover, ancestral sequences/structures determined in this project may be of interest for experimental work (see "Paleomolecular biochemistry", on page 364).

Other Future Work

One capability of MrBayes that we have not so far used[533] is examining the variation of rates at different positions. Correlations of this with structure may be of interest (Pollock & Bruno 2000).

[533] Note that this capability may be improved in the most recent version, with Gibbs sampling (Huelsenbeck *et al.* 2007).

Prediction without full modeling

Due to constraints such as volume packing[534], charges, hydrophobicity, and secondary structural tendencies[535], some amino acids will fit better into a given backbone geometry than will others (Aszodi, Munro, & Taylor 1997; Azarya-Sprinzak *et al.* 1997; Blundell 1991; Bowie, Luthy, & Eisenberg 1991; Fornasari, Parisi, & Echave 2002; Koehl & Levitt 2002; Ponder & Richards 1987a, 1987b; Sunyaev *et al.* 1997; Wilmanns & Eisenberg 1995). Thus, we should be able to determine how well each of our sets of possible amino acids (deduced from present-day protein sequences) for the next ancestral sequence down or up the tree fits the backbone geometry for a newly modeled structure. Those sequences that fit better should be considered more likely as the next stage, which along with other information (such as the likelihood of a given mutation) can be used to determine which sequences to model next. While this has been done in the present research to a minor degree (see "Usage of existing models", on page 343), it should be possible to do this on a larger scale (i.e., on more sequences) on a much more automated basis. Similar techniques may be useful in altering the substitution matrices according to the local residue environment (Fornasari, Parisi, & Echave 2002; Goldman, Thorne, & Jones 1998; Koshi & Goldstein 1995; Koshi, Mindell, & Goldstein 1997; Koshi & Goldstein 1998; Koshi, Mindell,

[534] I.e., the amount of space needed for a residue (Word *et al.* 2000).

[535] E.g., helix breaker versus helix former (Bowie, Luthy, & Eisenberg 1991; Wilmanns & Eisenberg 1995). One should keep in mind that some residues (e.g., proline (Gunasekaran *et al.* 1998; Prieto & Serrano 1997; Yang, W Z *et al.* 1998)) may act differently depending on where in the (potential) helix they are located (Cochran, Penel, & Doig 2001; Garnier, Osguthorpe, & Robson 1978; Kumar & Bansal 1998a, 1998b; Pace & Scholtz 1998).

& Goldstein 1999; Overington *et al.* 1992; Thompson, M J & Goldstein 1996; Topham *et al.* 1993; Wako & Blundell 1994a, 1994b).

Given the possibility of the local environment altering (Huang & Wang 2002), it may be advisable to do this in a "mixed" model similar to MrBayes' current multiple-matrix capability. Another (preferable if practical) possibility would be to have the possible secondary structure, *etc.* one of the characteristics evolving along the tree (Goldman, Thorne, & Jones 1996; Kawabata & Nishikawa 2000; Lio *et al.* 1998; Mizuguchi & Blundell 2000; Thorne, Goldman, & Jones 1996). The alignment database created as a part of the present research, given its inclusion of (generally high quality) structures, may be of use in determining such evolutionary patterns, and in the study of correlated mutations and their relationship to protein structure (see also "Sequence determination", on page 135).

Paleomolecular biochemistry

In the emerging field of paleomolecular biochemistry, the ancestral sequence of a protein is inferred, then that sequence is produced in the laboratory for further characterization (Adey *et al.* 1994; Chandrasekharan *et al.* 1996; Chang, B S W & Donoghue 2000; Dean & Golding 1997; Dean 1998; Golding & Dean 1998; Jermann *et al.* 1995; Miyazaki *et al.* 2001; Nei, Zhang, & Yokoyama 1997; Shi & Yokoyama 2003; Zhang, J & Rosenberg 2002; Zhang, J 2003). DHFR is eminently suitable for this (note requirement "D", on page 19); it:

- is well-characterized (see part 2, on page 49);
- has multiple known inhibitors with varying potency between species and sequences and available means of study (Appleman *et al.* 1988a; Appleman *et al.* 1990; Baccanari *et al.* 1989; Blakley & Sorrentino 1998; Brophy *et al.* 2000; Degan *et al.* 1989; Farnum *et al.* 1991; Lewis *et al.* 1995; Taira & Benkovic 1988);
- lacks characteristics that generally prevent study by X-ray crystallography[536] or NMR[537] (as can be seen by that structures of it done via both techniques are known); and
- can be characterized by means such as fluorescence (Appleman *et al.* 1988a; Appleman *et al.* 1990; Degan *et al.* 1989; Farnum *et al.* 1991; Rimet *et al.* 1987; Rimet *et al.* 1990; Rimet *et al.* 1991).

[536] E.g., prevent crystallization.
[537] E.g., stability with high salt, temperature, or protein concentration.

One area of future work, once:

- one or more possible sequences for the DHFR of the fungi/metazoa ancestor (or other points of interest) and their structural models have been determined to a higher degree of quality than currently; and
- a chain of models is successful in reaching and matching a modern DHFR structure among the fungi (*P. carinii* and/or *C. albicans*)

will be the production of the sequence or sequences and its/their characterization in the laboratory. Such characterization may ultimately include full structural determination via X-ray crystallography or NMR.

Appendix A: PDB files/chains used

Please see supplemental file “extract.important.pdbs.xls” or http://cesario.rutgers.edu/easmith/research/proteins/extract.important.pdbs.xls; the file is too large to include. We apologize for not citing at least the primary paper for each PDB file, but the number of PDB files involved (enough for 1938 chains) makes this impractical.

Please note that the last letter/digit in each name is the chain identifier, with "0" indicating no chain identifier unless other chains appear from that PDB file. Please also note that the chains used to designate each "cluster" (of 65%+ identical sequences), although not which chains are in each cluster, may differ from those used in further work. The assignment of clusters was by the program that interpreted the machine-readable form of this file (“extract.important.pdbs.txt”, also available as a supplemental file and via http://cesario.rutgers.edu/easmith/research/proteins/extract.important.pdbs.txt). The major program to interpret this file is "interpret.important.pdbs.pl"; the file is generated by "extract.important.pdbs.pl".

Appendix B: Important PDB files/chains used

Please see supplemental file “interpret.important.pdbs.xls”, also available as http://cesario.rutgers.edu/easmith/research/proteins/interpret.important.pdbs.xls; the table in question is too large to include directly (it has 312 chains from 278 PDB files[538]). Data values denoted by a "*" are estimates[539] or derived (e.g., the RMS) using mostly estimated data. For the meanings of the columns, the "Res" is resolution, the "Valid Length" is the number of amino acids in the PDB file with usable structures[540], and the "Length" is the number of residues in the PDB file's SEQRES. Please note that some chains are repeated multiple times; if the same (3D) *structural* sequence is found in more than one organism, then it is treated as being from multiple organisms. The PDB files listed were selected from those listed in Appendix A based on a combination of factors, including estimated RMS, Valid Length, and Length; these factors were also used in deciding on which PDB chain would be used for the name of the cluster (group of 65%+ identical chains). For a more UNIX-machine-readable version, please see the supplemental file "interpret.important.pdbs.txt.new.txt" (also available via http://cesario.rutgers.edu/easmith/research/proteins/interpret.important.pdbs.txt.new). This file

[538] Again, we apologize for not citing at least the primary paper for each PDB file, but the number of said files makes this impractical.
[539] For instance, if a Free R-value (R-free) was absent, it was estimated from the R-value by adding 0.086; this was the 95th percentile of the difference between R-values and Free R-values in the dataset. Note that the 95th percentile was used due to the suspicion that structures not providing Free R-values would be of lower quality due to probable overfitting, among other problems (Kleywegt & Jones 1995; Kleywegt & Brunger 1996).
[540] By "usable structures" is meant residues with backbone alpha carbon, N, O, C for all, and in addition beta carbon for non-glycine residues.

was generated by "interpret.important.pdbs.pl" (see "Appendix A: PDB files/chains used", on page 365).

Appendix C: Other sources for initial tree

A listing of TreeBASE (Sanderson *et al.* 1993) studies and trees from these studies used is in supplemental file "TreeBASE.trees.used.txt", also available at http://cesario.rutgers.edu/easmith/research/trees/TreeBASE.trees.used.txt. (In this file, study IDs (S###) are followed by a colon then a listing of the trees used.) We apologize that the number of trees involved makes (formally) citing each individual study impractical. Please also note that some (external) sources other than NCBI's taxonomy and TreeBASE were used (Angen *et al.* 2003; Chater & Horinouchi 2003; Ciccarelli *et al.* 2006; Embley & Stackebrandt 1994; Georis *et al.* 1999; Gerrits *et al.* 2005; Gophna, Doolittle, & Charlebois 2005; Gordon & Sibley 2005; Hankeln *et al.* 2006; Hoegger *et al.* 2006; Kampfer 2006; Kawase *et al.* 2004; Kim *et al.* 1999; Kirk *et al.* 2007; Kuhnert & Korczak 2006; Liu, Z *et al.* 2005; Matheny *et al.* 2007; Metsa-Ketela *et al.* 2002; Mouchacca 2000a, 2000b; Palleroni 2003; Ramachandra, Crawford, & Pometto 1987; Redfield *et al.* 2006; Romanelli, Houston, & Barnett 1975; Sproer *et al.* 1999; Stechmann & Cavalier-Smith 2003; Thompson, F L *et al.* 2005; Wang, S-J *et al.* 1999; Yumoto *et al.* 1998; Zrzavy 2001).

Appendix D: NCBI taxids and alternate species names

Listings of the following are available as the supplemental data files named in parentheses:

- species/subspecies names[541] versus NCBI id numbers[542] (species.names.NCBI.txt);
- equivalent data for genera and above (genus.above.names.NCBI.txt);
- species/subspecies correspondences (species.subspecies.NCBI.txt);
- lineages (species.lineage.NCBI.txt); and
- NCBI taxids not considered usable[543] (bad.nodes.txt).

The supplemental data files are also available via http://cesario.rutgers.edu/easmith/research/species/. Please see supplemental file "extract.species.used.taxdump.data.xls" for the portion of this data most important for the present work; it is also available at http://cesario.rutgers.edu/easmith/research/species/extract.species.used.taxdump.data.xls. The program for processing the NCBI taxonomy to produce the above files was “extract.species.names.pl”, with some manual editing with regard to important species (correcting the error of including the year from the “full” taxonomic name) and updates to the sequence databases (primarily for subspecies of species of importance).

---

[541] These include not only the locally used names but also others in the NCBI taxonomy database.
[542] These are as of the time of the downloading of the taxonomy files (Bischoff *et al.* 2004).
[543] For instance, NCBI’s taxids include identifiers for environmental samples; these were not considered usable.

Appendix E: MolProbity results

It has unfortunately not been possible to transfer the results (aside from the summary given under "8. Examination of models", on page 355) in any meaningful way to Microsoft Word format, due to their size. Please see supplemental file "extract.molprobity.1.new.xls" (also available online at http://cesario.rutgers.edu/easmith/research/molprobity/extract.molprobity.1.new.xls), which contains:

- A summary of the MolProbity results, in the "sheet" named "MolProbitySummary". This sheet also includes the filenames of the full MolProbity results. These files[544] are available online at http://cesario.rutgers.edu/easmith/research/molprobity/ and in the supplemental file “molprobity.html.tar”, which is a UNIX tar format archive.
- A listing of what residue sidechains/mainchains were considered "bad" for purposes of modeling (in the sheet named "MainSideBad").

Below is a table of the abbreviations used.

Abbrev.	Expansion	Explanation
freeze2	partially-frozen, vacuum	This refers to partially-frozen, vacuum/dry (without water) minimization; please see "Partially frozen vacuum/dry minimization", on page 167
full	full/wet, non-strict	This refers to full/wet (with water) minimization, with non-strict restraints (see "Creation of restraints", on page 170)
full-nadph	full/wet, only NADPH restrained, non-strict	This refers to full/wet (with water) minimization, with restraints placed on the NADPH only, with non-strict restraints
full-partial	full/wet, partial restraints, non-strict	This refers to full/wet (with water) minimization, with a partial set of non-strict restraints (see under "Creation of restraints", on page 173)

[544] We apologize for that these contain some JavaScript and/or other references to other pages; the functionality of the pages should not depend on these working.

Abbrev.	Expansion	Explanation
full-steep	full/wet, steep minimizer, non-strict	This refers to full/wet (with water) modeling, using the steepest descents minimizer (see "Full energy minimization", on page 181), with non-strict restraints
fullSA	full/wet, non-strict, SA	This refers to SA with water, with non-strict restraints. It is generally followed by a code indicating how the SA results were subsequently minimized (e.g., "full")
full2	full/wet, strict	This refers to full/wet modeling, with strict restraints
full2-nadph	full/wet, only NADPH restrained, strict	This refers to full/wet (with water) modeling, with restraints placed on the NADPH only, with strict restraints
full2-partial	full/wet, partial restraints, strict	This refers to full/wet (with water) minimization, with a partial set of strict restraints (see under "Creation of restraints", on page 173)
group1	group 1	For Uramniota, derived from the Urplacental stage output
group2	group 2	For Uramniota, derived from the two *G. gallus* (chicken) structures
Int. Stage	Intermediate Stage	At some points in the modeling, intermediate structures were constructed - these were not fully completed, but were far enough along for MolProbity evaluation to yield potentially-useful results in terms of what areas should be taken from which (intermediate) model
loop	Loop search	These were the products of a loop search for residues 13-27; see footnote 337, on page 157, for more information
mmtp	mmtp rotamer	Indicates the usage of the mmtp rotamer (Lovell *et al.* 2000) of lysine 127 for the initial structure
nH	new Hydrogens	A corrected version of translation of hydrogens in NADPH between GROMACS' and `reduce`'s formats
nonrotamer	non rotamer library	Indicates the usage of a non-rotamer library rotamer (found by `probe` (Word *et al.* 2000)) for lysine 127 for the initial structure
Rama.	Ramachandran	Backbone angles (phi/psi)
vacuum2	2nd vacuum minimization	After the 2nd stage of vacuum minimization; please see "Non-frozen vacuum minimization", on page 174
vacuum3	3rd vacuum minimization	After the 3rd stage of vacuum minimization; please see under "Non-frozen vacuum minimization", footnote 375, on page 174
Y-rotamer	Rotamer search for tyrosine 177	Please see "Rotamer searches", on page 371
YS-rotamer	Rotamer search for tyrosine, serine	Please see "Rotamer searches", on page 371

Appendix F: Proteins removed

The following proteins were not used in tree determination at or after the stage of predicting the fungi/metazoa ancestral DHFR sequence, due to problems with the *Neurospora crassa* DHFR sequence:

- Cellulase B (glycosyl hydrolase 6; 1,4-beta-cellobiohydrolase) (Coutinho, P M & Henrissat 1999; Coutinho, Pedro M & Henrissat 2007)
- Cellulase C (glycosyl hydrolase 7; 1,4-beta-cellobiosidase/endo-1,4-beta-glucanase) (Coutinho, P M & Henrissat 1999; Coutinho, Pedro M & Henrissat 2007)
- Cellulase F (glycosyl hydrolase 10; endo-1,4-beta-xylanase) (Coutinho, P M & Henrissat 1999; Coutinho, Pedro M & Henrissat 2007)
- Cellulase G (glycosyl hydrolase 11; endo-1,4-beta-xylanase; E.C. 3.2.1.8) (Coutinho, P M & Henrissat 1999; Coutinho, Pedro M & Henrissat 2007)

Appendix G: ESIMILARITY matrix

	V	I	L	M	W	F	Y	T	A	C	G	P	R	S	H	N	K	Q	E	D	Z	B
V	**1**	*1*	*1*	*1*	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
I	*1*	***1***	*1*	*1*	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
L	*1*	*1*	**1**	*1*	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
M	*1*	*1*	*1*	**1**	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
W	-1	-1	-1	-1	**1**	*1*	*1*	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
F	-1	0	0	0	*1*	**1**	*1*	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Y	-1	-1	-1	-1	*1*	*1*	**1**	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1
T	-1	-1	-1	-1	-1	-1	-1	**1**	-1	-1	-1	-1	-1	*1*	-1	0	-1	-1	-1	-1	-1	-1
A	-1	-1	-1	-1	-1	-1	-1	-1	**1**	0	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1
C	-1	-1	-1	-1	-1	-1	-1	-1	0	**1**	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
G	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	**1**	-1	-1	0	-1	0	-1	-1	-1	-1	-1	-1
P	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	**1**	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
R	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	**1**	-1	0	0	*1*	*1*	0	-1	0	-1
S	-1	-1	-1	-1	-1	-1	-1	*1*	0	-1	0	-1	-1	**1**	-1	*1*	0	0	0	0	0	0
H	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	0	-1	**1**	*1*	-1	0	0	-1	0	0
N	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	0	-1	0	*1*	*1*	**1**	0	0	0	*1*	0	*1*
K	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	*1*	0	-1	0	**1**	*1*	*1*	-1	*1*	0
Q	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	*1*	0	0	0	*1*	**1**	*1*	0	*1*	0
E	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	0	*1*	*1*	**1**	*1*	*1*	*1*
D	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	*1*	-1	0	*1*	**1**	*1*	*1*
Z	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	0	*1*	*1*	*1*	*1*	**1**	*1*
B	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	*1*	0	0	*1*	*1*	*1*	**1**

The above matrix (for usage in sequence comparisons only, not for searching[545] or alignment) was created (by the program "create.similarity.matrix.1.pl") via a combination[546] of the BLOSUM62 (Henikoff & Henikoff 1992), “Nussinov” (Naor *et al.* 1996), and Ident matrices. (It includes B (N or D) and Z (Q or E) because some sequences from SWISS-PROT (Boeckmann *et al.* 2003) included these ambiguity codes.) For ease of interpretation, the diagonal entries are in **bold**, and other positive entries are in *italics*.

[545] At least, not for direct use in searching; it was used in the construction of groups for scanprosite (see “Loop searches”, on page 157).

[546] The combination requires 2 of 3 of the matrices to be positive for the result to be positive; please see the program code for more details as to the handling of, for instance, entries of "0".

Appendix H: Evaluation of alignment quality

Initially, clustered (into 65%+ identity groups) PDB chains were aligned by the programs "align.clustered.pdbs.pl", "align.clustered.pdbs.2.pl", and "align.clustered.pdbs.3.pl". These were analyzed using other programs[547], such as “analyze.align.clustered.pdbs.9.pl” and “analyze.align.clustered.pdbs.11.pl”. The program "find.align.thresholds.pl" used the output from these to derive thresholds (for minimum quality, according to the proportion aligned and percent identity), along with manual intervention when the automated procedure failed due to overlaps (particularly in the larger dataset derived from analysis using "analyze.align.clustered.pdbs.11.pl"). The results from "analyze.align.clustered.pdbs.9.pl" (in ".csv" format) were analyzed using a nonlinear least squares equation solver to derive an equation for deciding on the quality within groups satisfying the thresholds. Also derived from the above programs were listings of what matrices appeared to work best for each cluster. All of the datafiles in question are available at http://cesario.rutgers.edu/easmith/research/analyze.align/.

[547] The programs used were analyze.align.clustered.pdbs.2.pl, analyze.align.clustered.pdbs.3.pl, analyze.align.clustered.pdbs.4.pl, analyze.align.clustered.pdbs.5.pl, analyze.align.clustered.pdbs.6.pl, analyze.align.clustered.pdbs.9.pl, analyze.align.clustered.pdbs.10.pl, and analyze.align.clustered.pdbs.11.pl.

Appendix I: Species groupings used

The following were the species groupings used for ambiguity simplification (see "Further sequence processing: Ambiguity-coded polymorphism reduction", on page 94), group sequence creation (see "Further sequence processing: Group sequence creation", on page 96), and sometimes constraints on tree searches (see “Tree searches”, on page 299), including at later stages:

- Archaea
- Bacteria
- Proteobacteria
- Eukaryota
- Fungi
- Metazoa
- Fungi/Metazoa
- Possible Fungi/Metazoa; this grouping includes, as well as fungi/metazoa:
  - *D. discoideum*
  - *E. histolytica*
  - *Hartmannella cantabrigiensis*

  Note that some initial tree search runs were done without any constraint on *D. discoideum* and *E. histolytica* being together with Fungi/Metazoa (*Hartmannella cantabrigiensis* was not added until after some other DHFR sequences were incorporated). These runs did not result in moving them away, although said runs were sufficiently problematic in other respects

(based on comparisons with phylogenies accepted by anyone in the field) that this is not much of an argument. Note, however, that neither *D. discoideum* nor *E. histolytica* have (used/known) DHFR sequences, decreasing their importance for this work, and that *Hartmannella cantabrigiensis*'s position is supported by prior research (Stechmann & Cavalier-Smith 2003) with no other research, as far as we are aware, contradicting it.

- Tetrapoda
- Vertebrata
- Mammalia
- Aves (birds) - this grouping was used only for ambiguity reduction, not for group sequence creation, due to its low number of species (in our database)
- Plant/Algae (Viridiplantae)

Please see supplemental file "species.groups.txt"[548] for information on what species are in each group; this file's format is a listing of "constraints" to be used for MrBayes, and do include "full" species (see "Usage of polymorphism", on page 64). Some additional, possibly non-clade[549] groupings were also used, although these were avoided when possible after the DHFR sequences were input into the alignment:

- Bacteria other than Proteobacteria
- Eukaryota other than Fungi/Metazoa
- Metazoa other than Vertebrata

[548] This can also be found at http://cesario.rutgers.edu/easmith/research/species/species.groups.txt.
[549] A clade can be defined as a group of species more closely descended from a common ancestor than the other species being considered in a study (Futuyma 1986).

- Vertebrata other than Tetrapoda
- Tetrapoda other than Mammalia

The following were used only for earlier stages (e.g., when not using "create.outgroup.seqs.pl" - see "Further sequence processing: Group sequence creation", on page 96):

- Possible Fungi/Metazoa other than known Metazoa
- Possible Fungi/Metazoa other than known Fungi

A few other groups (e.g., Alveolata) have been used for presentation purposes but not for tree work; these are named according to the NCBI taxonomy insofar as it agrees with this work’s findings.

Appendix J: MrBayes review/explanation

MCMC

In MrBayes' "MCMC" (Monte Carlo Markov Chain), an initial model, including a tree topology with branch lengths, is initially generated (randomly for some parameters, from input information for others - see "Usage of the results of prior tree runs", on page 127). The model is then altered, with random elements (thus "Monte Carlo"), by the "moves" (see below). (The model is not newly generated each time, but is a modified version of the previous model, as indicated by the name "Markov Chain".) Samples are taken at regular intervals (every 100 moves tried ("generations"), as per the MrBayes default). A subset of these samples is then used by the "sumt" and "sump" commands, with the portion not used being determined by the "burnin" (see footnote 428 under "Simulated Annealing (SA)", on page 197).

Short summary of moves

MrBayes functions via applying "moves" (also known as "proposals") to alter various aspects of the modeled tree (ranging from the topology of the tree, to the distances on the tree, to the rate variations along the sequences (gamma)). These moves are applied on a random basis; their result is then evaluated for how probable the resulting tree would be, and accepted or rejected on this basis (and on the basis of how probable the move in question is - a more extreme move is less likely to be accepted). To be noted is that even a move that

decreases the likelihood of the tree may be accepted, on a chance basis, particularly if the "temperature" is increased, so that the algorithm is not trapped into a local minimum.

More detailed description/explanation of moves

The likelihood of a “move” being accepted[550] is dependent on both:

1. How the likelihood of the tree changes; and
2. How likely the given “move” is to be acceptable in the first place[551].

The “temperature” affects the first of these; the proposal settings (see page 381) affect both. If the above indicate that the result (the combination of the two) is unlikely, then there is a chance that the move will nevertheless be accepted; if the above indicates that the result is likely, then the move will always be accepted. This is a “Metropolis-Hastings” (Hastings 1970; Metropolis *et al.* 1953) algorithm. The likelihood of a “wrong-way” move[552] is determined by a Boltzmann (Wikipedia 2008) distribution (with a temperature[553] above absolute zero). It is

[550] This procedure may seem overly elaborate. However, it has been found to be easier to first propose a possible move then see whether it is acceptable than to generate a known-acceptable move in the first place; if the latter was possible, then there would be a closed-form solution to the problem.

[551] The latter consideration is, in many cases, at least partially in order to make the proposal and parameter distribution an appropriate one - e.g., to convert it from a uniform distribution to a lognormal or normal one.

[552] Why are “wrong-way” moves sometimes wanted? Moving to a less-likely situation makes possible a further, perhaps a more radical change that would not be accessible if only moves to more likely situations were accepted. Again, it is similar to a way out of a potential energy well, which may be desirable if there is a deeper well - ideally, the deepest well, the goal - “someplace” else.

[553] In the “classical” form, the “temperature” used by MrBayes would simply have been this temperature. However, since it is desired to have the temperature only affect which moves are accepted due to tree changes, not due to how extreme the moves are (the latter should be set by the proposal parameters) without regard to the effects on the tree, the “temperature” used instead affects the process prior to this. It does this by decreasing the weight of the log probabilities for everything but the move’s “prior” probability, prior to the addition of the log of the move’s prior probability. It is thus done as a multiplier by a number below 1 to yield a *higher* temperature in

equivalent to the likelihood of a particle jumping to a higher energy state - e.g., out of a potential energy well (a local minimum).

The settings of a move/proposal determine what degrees of changes can be tried. For most moves, this is either a ± modification[554] to the existing parameter, for a "sliding window", or a multiplicative version of this (essentially a "sliding window" on a log scale); see under "Adapt", on page 382. Other moves, such as tree rearrangements and adjustments of estimated state frequencies (since the latter have to add up to 100%; these are the "Dirichlet" state frequencies - see "Partitions: State frequencies", on page 107).

Move acceptance percentages

It is generally recommended for MrBayes (Huelsenbeck *et al.* 2006; Ronquist 2005) that the percentage of moves accepted be in the ~10/20-70% range. If the percentage of accepted moves is below 10/20%, then the program is spending too much time trying moves that do not work. If the percentage of moves accepted is above 70%, then the program is not trying large enough moves to get out of any local minima.

Adapt and SA

Both Adapt and SA (see "MrBayes code alterations", on page 98) function to try to keep the percentage of moves accepted within the above limits. Larger

terms of the original algorithm - see footnote 202 under "MrBayes code alterations", on page 100.
[554] I.e., a range of possibilities above and below the existing parameter.

changes are, both in terms of the likelihood of the tree and (for some cases) the likelihood of the move, less likely to be acceptable. However, while small changes are less likely to cause problems, they are also not as likely to jump out of a local minimum (a local potential energy well).

Adapt

Adapt adjusts the settings of the move. It is only able to act on “sliding window” and “multiplier” moves. The former type of move adds or subtracts a random number - with a maximum of the setting, namely the “sliding window” size - from the current value of a given parameter of the model. For instance, if the current value of the parameter was 0.2, and the sliding window size was 0.1, then the move could result in a parameter from 0.1 to 0.3 - this is a window on a number line, which slides up and down with the current value of the parameter. Multiplier moves instead multiply or divide by a random number, again with a maximum multiplier/divider determined by the setting for the move. This action is equivalent to a sliding window on a log scale. Adapt works by adjusting the settings upward (allowing larger moves) if there are too many acceptances, and vice-versa if there are too few.

SA

SA adjusts the “temperature” of the move. In a physical simulation, this would be adjusting the likelihood that a particle would have enough energy to jump out of a local minimum (or to jump out of the *correct* minimum, unfortunately). It adjusts

the temperature at the beginning of the run to allow for more freedom, then adjusts it back to the normal setting at a rate determined by how often moves are being accepted/refused. It is hoped that in the initial portion of the run, the moves will have ranged far enough to find the correct “region”, while in the later portions, the smaller moves will home in on the best “location” within that “region” without jumping too far away.

Adapt, SA, and burnin

Both Adapt and SA use the “burnin” value for the “mcmc” command (which initiates a run) to avoid doing any changes in the later portion of a run (which will be used for sampling). This burnin value, when Adapt and/or SA are used, is therefore set to the minimum burnin expected to be used for the “sump” or “sumt” commands (see footnote 428 under “Simulated Annealing (SA)”, on page 197).

Appendix K: Partial DHFR alignment

It has unfortunately not been possible to give the entire alignment in a readable format in the body of the dissertation; see "5. Alignment of central sequences", on page 336. This appendix contains a partial version. The below table gives the identity of the species associated with the "Id" involved, as given in the alignment.

Id	Species	Common/other name
1U70A	*Mus musculus*	Mouse
1U72A	*Homo sapiens*	Human
8DFR0	*G. gallus*	Chicken
XP_001176553	*Strongylocentrotus purpuratus*	Sea urchin
DYR_DROME	*Drosophila melanogaster*	Fruit fly
EAL28532	*Drosophila pseudoobscura*	Fruit fly
Q7Q0L5_ANOGA	*Anopheles gambiae*	Mosquito
XP_393902	*Apis mellifera*	Honeybee
XP_973338	*Tribolium castaneum*	Flour beetle
DYR_CAEEL	*C. elegans*	(Nematode[555])
Q61DT5_CAEBR	*C. briggsae*	(Nematode)
DYR_ENCCU	*Encephalitozoon cuniculi*	(Microsporidia)
BAC75955	*Coprinus cinereus*	(Basidiomycota)
DYR_CRYNE	*Cryptococcus neoformans*	(Basidiomycota)
EAK84413[556]	*Ustilago maydis*	(Basidiomycota)
DYR_PNECA	*P. carinii*[557]	(Ascomycota[558])
AAF14071		
DYR_SCHPO	*S. pombe*	(Ascomycota)
DYR_CANAL	*C. albicans*	(Ascomycota)
CAG60823	*Candida glabrata*	(Ascomycota)
DYR_YEAST	*S. cerevisiae*	(Ascomycota)
1J3IA	*P. falciparum*	Malaria
2BL9A	*P. vivax*	Malaria
1SEJC	*Cryptosporidium hominis* *Cryptosporidium parvum*	

[555] A microscopic type of worm; *C. elegans* is among the model species most used in biology.
[556] Note that the last few characters of this sequence have been removed; they did not correspond to any others as far as could be told, and resulted in an extra page of the table with only two lines of sequence. (The sequence in question is "tv".)
[557] These are from the rat and (a major strain of) human *P. carinii*, respectively; the target structure is of the rat sequence (DYR_PNECA).
[558] Ascomycota are the group of fungi including most studied yeasts.

In the alignment (see following pages), note that "O. Amniota" stands for "Other Amniota", as in Amniota (e.g., birds and marsupials) other than placental mammals. Similarly, "O. Deuterostomia" stands for "Other Deuterostomia" (ones other than Amniota), and "O. Fungi" stands for "Other Fungi" (ones other than Ascomycota). Predicted ancestral sequences are not in the above table; they are given in the alignment under the grouping to which they are ancestral (e.g., Uramniota is in the section with “Other Amniota”); see under “6. Determination of ancestral sequences”, on page 134, for more on the meaning of the names. If an "Id" is over 2 columns, then it has been entered with two different alignments due to uncertainty (see under "Alignment using HMM", on page 131). The "Type" is "S" for a structure ("S*" for structures not used, due to being targets), M for models ("M*" if a model will not be used further, and "M(*)" if a model would normally be used further but for the problems noted in "8. Examination of models", on page 352). The position in the alignment is indicated by the column of numbers on one side of the table.

Areas of the alignment with at least some letters in uppercase are those that were considered structurally alignable (neither “uncertain” nor “nonstruct”). We apologize for any difficulty in reading the below table, and note that it is also available as the supplemental file "DHFR.with.fungi.2.seqs.edited.7.vert.xls" and at http://cesario.rutgers.edu/easmith/research/DHFR.with.fungi.2.seqs.edited.7.vert.xls for viewing with better zoom capabilities than those supplied by a magnifying glass.

Group:	Placental			O. Amniota			O. Deuterostomia					Insecta					Nematoda				Fungi/Metazoa								O. Fungi					Ascomycota																										Alveolata		
Id:	1U70A	1U72A	Urplacental	8DFR0	AGA	PGV	XP_001176553	ES	KAG	K_D	K_K	DYR_DROME	EAL28532	Q7Q0L5_ANOGA	XP_393902	XP_973338	DYR_CAEEL		Q61DT5_CAEBR		1111_GTF	1111_GVF	1111_STF	1111_SVF	1111_chars2	1100_SVFQ	1100_SVYQ	1100_chars2	DYR_ENCCU		BAC75955	DYR_CRYNE	EAK84413	0010_KD	0010_KP	0010_QD	0010_QP	0011_KD	0011_KP	0011_QD	0011_QP	10100_chars2	10101_chars2	10110_chars2	10111_chars2	11100_chars2	11101_chars2	11110_chars2	11111_chars2	DYR_PNECA	AAF14071	DYR_SCHPO		DYR_CANAL		CAG60823		DYR_YEAST		1J3IA	2BL9A	1SEJC
Type:	S	S	M	S	M	M		M	M	M*	M*										M	M*	M	M*	M*	M*	M*	M*						M*	M*	M*	M*	M(*)	M(*)	M(*)	M(*)	M*	M*	M*	M*	M*	M*	M*	M*	S*				S*	S*					S	S	S
386	V	V	V	V	V	V	I	V	V	v	v	I	V	I	V	V	K	K	K	K	M	m	M	m	v	m	m	v	L	L	M	M	R	M	M	M	M	M	M	M	M	m	m	m	m	m	m	m	m	M	M	L	L	L	L	L	L	L	L	I	V	I
387	Y	Y	Y	Y	Y	Y	Y	Y	Y	y	y	L	L	Y	Y	Y	W	W	W	W	Y	y	Y	y	y	y	y	y	Y	Y	W	W	L	W	W	W	W	W	W	W	W	w	w	w	w	w	w	w	w	W	W	Y	Y	W	W	Y	Y	Y	Y	Y	Y	F
388	E	E	E	Q	E	E	E	E	E	e	e	E	E	E	K	Q	K	K	K	K	T	t	T	t	t	t	t	t	T	T	T	V	P	T	T	T	T	T	T	T	T	t	t	t	t	t	t	t	t	T	I	G	G	T	T	T	T	N	N	K	S	E
389	K	K	K	K	K	K	S	K	K	k	k	K	K	N	K	N	V	V	I	I	R	r	R	r	k	r	r	k	K	K	R	L	N	R	R	R	R	R	R	R	R	r	r	r	r	r	r	r	r	R	R	K	K	R	R	K	K	R	R	K	K	K
390	k	n	n	s	n	n	s									p					s	s	s	s	s	s	s	s	p	p	s	n		S	S	S	S	S	S	S	S	x	x	x	x	x	x	x	x	d	d	n	n	k	k	k	k	k	k	t		q
391				v													v	v	i	i													l	i	i	i	i	i	i	i	i	i	i	i	i	i	i	i	i	l	i										v	
392	d	d	d		e	e									e	d	e	e	e	e													l																											n		e
393																	n	n													s		k																			d	d							n		k
394														k																		q	q																			d	d							k		k
395				l																																																v	v							m	g	t
396				a																													q																													l
397				q								h					h	h	h	h									e	e			e																			n	n									q
398																f																	w																			v	v									
399							s					s				n													t	t			t																			a	a									
400																													r	r			q																													
401																																	a																			l	l								g	
402																																	s																													
403																																	a																				l									
404																																	d																				e									
405																																	s																													
406																																	l																			l										
407																																	r																			e	k									
408																																	q																			k										
409																																	y																													
410																																	l																			l	l									
411																																	g																													
412																																	s																			s	s									
413																																	a																			i	i									
414																			c	c													c																			c	c									
415																																	p																													
416																																	v																													
417																																	a																													
418																																	l																													
419																																	t																													
420																																	d																													
421																																	s																													
422																																	r																													
423																																	d																													
424																																	m																													
425																																	v																													
426																																	t																													
427																																	s																													
428																																	e																													
429																																	g																													
430																																	e																													
431																																	t																													
432																																	w																													
433																																	y																													
434																																	q																													
435																																	y																													
436																																	q																													
437																																	l																													
438																																	w																													
439																																	e																													
440																																	k																													

Appendix L: Tree files available, cross-referenced to pictures

The following PHYLIP-format (Felsenstein 1993) tree files are available in the supplemental data file “trees.tar”, which is in UNIX “tar” format, and under http://cesario.rutgers.edu/easmith/research/trees/. Noted are figures in this text in which they are pictured; on page 400 is a table cross-referencing page numbers with tree figures to the files from which they are derived.

- DHFR.alveolata.kinetoplastida.plants.phy
  - Figure 4.T.nfm, on page 328
- DHFR.fungi.phy
  - Figure 4.T.fungi.p, on page 329
  - Figure 4.T.fungi.c, on page 330
- DHFR.invertebrates.phy
  - Figure 4.T.invertebrates, on page 331
- DHFR.real.structures.phy
  - Figure 3.1, on page 52
- DHFR.sequences.example.phy
  - Figure 1.1, on page 4
- DHFR.structures.partial.2.phy
  - Figure 3.4, on page 149
- DHFR.structures.partial.phy
- DHFR.vertebrata.phy
  - Figure 4.T.vertebrata, on page 332
- archaea.phy
- bacteria_non_proteobacteria.phy

- eukaryota.tree.search.phy
  - Figure 4.T.s.eukaryota.p, on page 301
  - Figure 4.T.s.eukaryota.c, on page 302
- fungi.phy
- original.round1.phy
  - Figure 4.1, on page 193
- primates.rodentia.tree.search.phy
  - Figure 4.T.s.mammalia.p, on page 317
  - Figure 4.T.s.mammalia.p.tetrapoda, on page 318
  - Figure 4.T.s.mammalia.c, on page 319
- proteobacteria.phy
- proteobacteria.tree.search.phy
  - Figure 4.T.s.proteobact.p:, on page 306
  - Figure 4.T.s.proteobact.p.proteobact, on page 307
  - Figure 4.T.s.proteobact.c, on page 308
- round1.subset.1.current.phy:
  - Figure 4.T.r1.s1.c.p, on page 237
  - Figure 4.T.r1.s1.c.p.proteobacteria, on page 238
  - Figure 4.T.r1.s1.c.p.eukaryota, on page 239
  - Figure 4.T.r1.s1.c.c, on page 240
- round1.subset.1.orig.phy
- round1.subset.1.usertree.12.phy
- round1.subset.1.usertree.13.phy
- round1.subset.2.current.phy:
  - Figure 4.T.r1.s2.c.p, on page 208
  - Figure 4.T.r1.s2.c.p.eukaryota, on page 209
  - Figure 4.T.r1.s2.c.p.bacteria, on page 210
  - Figure 4.T.r1.s2.c.c, on page 211

- round1.subset.2.orig.phy
  - Figure 4.T.r1.s2.1, on page 212
  - Figure 4.T.r1.s2.1.eukaryota, on page 213
  - Figure 4.T.r1.s2.1.bacteria, on page 214
- round1.subset.2.usertree.12.phy
  - Figure 4.T.r1.s2.12, on page 215
  - Figure 4.T.r1.s2.12.eukaryota, on page 216
- round1.subset.2.usertree.13.phy
  - Figure 4.T.r1.s2.13, on page 217
  - Figure 4.T.r1.s2.13.eukaryota, on page 218
- round1.subset.2.usertree.15.phy
  - Figure 4.T.r1.s2.15, on page 219
  - Figure 4.T.r1.s2.15.bacteria, on page 220
- round1.subset.3.current.phy
  - Figure 4.T.r1.s3.c.p, on page 254
  - Figure 4.T.r1.s3.c.p.eukaryota, on page 255
  - Figure 4.T.r1.s3.c.c, on page 256
- round1.subset.3.orig.phy
  - Figure 4.T.r1.s3.1, on page 257
  - Figure 4.T.r1.s3.1.saccharomycotina, on page 258
- round1.subset.3.usertree.5.phy
  - Figure 4.T.r1.s3.5, on page 259
  - Figure 4.T.r1.s3.5.saccharomycotina, on page 260
- round1.subset.3.usertree.6.phy
  - Figure 4.T.r1.s3.6, on page 261
  - Figure 4.T.r1.s3.6.saccharomycotina, on page 262
- round1.subset.4.current.phy
- round1.subset.4.orig.phy

- round1.subset.4.usertree.7.phy
- round1.subset.5.current.phy
  - Figure 4.T.r1.s5.c.p, on page 223
  - Figure 4.T.r1.s5.c.p.eukaryota, on page 224
  - Figure 4.T.r1.s5.c.c, on page 225
- round1.subset.5.orig.phy
  - Figure 4.T.r1.s7.1, on page 246
- round1.subset.5.usertree.14.phy
  - This tree moves one species, *Ommastrephes sloani*, the (unfortunately) sole Mollusca in the dataset. As might be expected (in hindsight), there was no significant difference between the log probabilities for it and the original tree (data not shown).
- round1.subset.5.usertree.2.phy
- round1.subset.5.usertree.3.phy
- round1.subset.5.usertree.4.phy
- round1.subset.6.current.phy
  - Figure 4.T.r1.s6.c.p, on page 232
  - Figure 4.T.r1.s6.c.p.eukaryota, on page 233
  - Figure 4.T.r1.s6.c.c, on page 234
- round1.subset.6.orig.phy
  - Figure 4.T.r1.s6.1, on page 235
- round1.subset.6.usertree.12.phy
- round1.subset.6.usertree.13.phy
- round1.subset.6.usertree.2.phy
- round1.subset.6.usertree.3.phy
- round1.subset.6.usertree.4.phy

- round1.subset.7.current.phy
  - Figure 4.T.r1.s7.c.p, on page 243
  - Figure 4.T.r1.s7.c.p.eukaryota, on page 244
  - Figure 4.T.r1.s7.c.c, on page 245
- round1.subset.7.orig.phy
  - Figure 4.T.r1.s7.1, on page 246
  - Figure 4.T.r1.s7.1.saccharomycotina, on page 247
- round1.subset.7.usertree.12.phy
- round1.subset.7.usertree.13.phy
- round1.subset.7.usertree.5.phy
  - Figure 4.T.r1.s7.5, on page 248
  - Figure 4.T.r1.s7.5.saccharomycotina, on page 249
- round1.subset.7.usertree.6.phy
  - Figure 4.T.r1.s7.6, on page 250
  - Figure 4.T.r1.s7.6.saccharomycotina, on page 251
- round1.subset.8.current.phy
- round1.subset.8.orig.phy
- round1.subset.8.usertree.2.phy
- round1.subset.8.usertree.3.phy
- round1.subset.8.usertree.4.phy
- round2.subset.10.current.phy
  - Figure 4.T.r2.s10.c.p, on page 285
  - Figure 4.T.r2.s10.c.p.eukaryota, on page 286
  - Figure 4.T.r2.s10.c.c, on page 287
- round2.subset.10.orig.phy
  - Figure 4.T.r2.s10.1, on page 288
  - Figure 4.T.r2.s10.1.nfm, on page 289
- round2.subset.10.usertree.11.phy

- round2.subset.10.usertree.12.phy
- round2.subset.10.usertree.2.phy
  - Figure 4.T.r2.s10.2, on page 290
- round2.subset.10.usertree.3.phy
  - Figure 4.T.r2.s10.3, on page 291
- round2.subset.12.current.phy
  - Figure 4.T.r2.s12.c.p, on page 293
  - Figure 4.T.r2.s12.c.p.eukaryota, on page 294
  - Figure 4.T.r2.s12.c.c, on page 295
- round2.subset.12.orig.phy
  - Figure 4.T.r2.s12.1, on page 296
- round2.subset.12.usertree.11.phy
- round2.subset.12.usertree.12.phy
- round2.subset.12.usertree.5.phy
  - Figure 4.T.r2.s12.5, on page 297
- round2.subset.8.current.phy
  - Figure 4.T.r2.s8.c.p, on page 269
  - Figure 4.T.r2.s8.c.p.eukaryota, on page 270
  - Figure 4.r2.s8.c.c, on page 271
- round2.subset.8.orig.phy
  - Figure 4.T.r2.s8.1, on page 272
  - Figure 4.T.r2.s8.1.mammalia, on page 273
  - Figure 4.T.r2.s8.1.nfm, on page 274
- round2.subset.8.usertree.10.phy
  - Figure 4.T.r2.s8.10, on page 278
  - Figure 4.T.r2.s8.10.mammalia, on page 279

- round2.subset.8.usertree.11.phy
  - Figure 4.T.r2.s8.11, on page 280
  - Figure 4.T.r2.s8.11.nfm, on page 281
- round2.subset.8.usertree.12.phy
  - Figure 4.T.r2.s8.12, on page 282
  - Figure 4.T.r2.s8.12.nfm, on page 283
- round2.subset.8.usertree.2.phy
  - Figure 4.T.r2.s8.2, on page 275
- round2.subset.8.usertree.9.phy
  - Figure 4.T.r2.s8.9, on page 276
  - Figure 4.T.r2.s8.9.mammalia, on page 277
- round4.subset.10.tree.search.phy
  - Figure 4.T.s.nfm.p, on page 314
  - Figure 4.T.s.nfm.p.eukaryota, on page 315
- round4.tree.search.subset.15.phy
  - Figure 4.T.s.insecta.p, on page 310
  - Figure 4.T.s.insecta.p.metazoa, on page 311
  - Figure 4.T.s.insecta.c, on page 312
- round7.subset.15.current.phy
  - Figure 4.T.r7.s15.c.p, on page 322
  - Figure 4.T.r7.s15.c.p.eukaryota, on page 323
  - Figure 4.T.r7.s15.c.p.fungi, on page 324
  - Figure 4.T.r7.s15.c.c, on page 325
- round7.subset.15.orig.phy
  - Figure 4.T.s.r7.s15.1, on page 326

Figure	Page	File
Figure 1.1	4	DHFR.sequences.example.phy
Figure 3.1	52	DHFR.real.structures.phy

Figure	Page	File
Figure 3.4	149	DHFR.structures.partial.2.phy
Figure 4.1	193	original.round1.phy
Figure 4.T.r1.s2.1	212	round1.subset.2.orig.phy
Figure 4.T.r1.s2.1.eukaryota	213	
Figure 4.T.r1.s2.1.bacteria	214	
Figure 4.T.r1.s2.12	215	round1.subset.2.usertree.12.phy
Figure 4.T.r1.s2.12.eukaryota	216	
Figure 4.T.r1.s2.13	217	round1.subset.2.usertree.13.phy
Figure 4.T.r1.s2.13.eukaryota	218	
Figure 4.T.r1.s2.15	219	round1.subset.2.usertree.15.phy
Figure 4.T.r1.s2.15.bacteria	220	
Figure 4.T.r1.s5.c.p	223	round1.subset.5.current.phy
Figure 4.T.r1.s5.c.p.eukaryota	224	
Figure 4.T.r1.s5.c.c	225	
Figure 4.T.r1.s5.1	226	round1.subset.5.orig.phy
Figure 4.T.r1.s6.c.p	232	round1.subset.6.current.phy
Figure 4.T.r1.s6.c.p.eukaryota	233	
Figure 4.T.r1.s6.c.c	234	
Figure 4.T.r1.s6.1	235	round1.subset.6.orig.phy
Figure 4.T.r1.s1.c.p	237	round1.subset.1.current.phy
Figure 4.T.r1.s1.c.p.proteobacteria	238	
Figure 4.T.r1.s1.c.p.eukaryota	239	
Figure 4.T.r1.s1.c.c	240	
Figure 4.T.r1.s7.c.p	243	round1.subset.7.current.phy
Figure 4.T.r1.s7.c.p.eukaryota	244	
Figure 4.T.r1.s7.c.c	245	
Figure 4.T.r1.s7.1	246	round1.subset.7.orig.phy
Figure 4.T.r1.s7.1.saccharomycotina	247	
Figure 4.T.r1.s7.5	248	round1.subset.7.usertree.5.phy
Figure 4.T.r1.s7.5.saccharomycotina	249	
Figure 4.T.r1.s7.6	250	round1.subset.7.usertree.6.phy
Figure 4.T.r1.s7.6.saccharomycotina	251	
Figure 4.T.r1.s3.c.p	254	round1.subset.3.current.phy
Figure 4.T.r1.s3.c.p.eukaryota	255	
Figure 4.T.r1.s3.c.c	256	
Figure 4.T.r1.s3.1	257	round1.subset.3.orig.phy
Figure 4.T.r1.s3.1.saccharomycotina	258	
Figure 4.T.r1.s3.5	259	round1.subset.3.usertree.5.phy
Figure 4.T.r1.s3.5.saccharomycotina	260	
Figure 4.T.r1.s3.6	261	round1.subset.3.usertree.6.phy
Figure 4.T.r1.s3.6.saccharomycotina	262	
Figure 4.T.r2.s8.c.p	269	round2.subset.8.current.phy
Figure 4.T.r2.s8.c.p.eukaryota	270	
Figure 4.r2.s8.c.c	271	
Figure 4.T.r2.s8.1	272	round2.subset.8.orig.phy
Figure 4.T.r2.s8.1.mammalia	273	
Figure 4.T.r2.s8.1.nfm	274	
Figure 4.T.r2.s8.2	275	round2.subset.8.usertree.2.phy
Figure 4.T.r2.s8.9	276	round2.subset.8.usertree.9.phy
Figure 4.T.r2.s8.9.mammalia	277	
Figure 4.T.r2.s8.10	278	round2.subset.8.usertree.10.phy
Figure 4.T.r2.s8.10.mammalia	279	

Appendix M: Model PDB-format files

The following PDB-format files are available[559] in the supplemental data file "struct.tar" (in UNIX "tar" format) and under http://cesario.rutgers.edu/easmith/research/struct/:

- 1U70A.f2p.nowat.reduce3.ent
- 1U70A.fp.nowat.reduce3.ent
- 1U72A.NA.f2p.nowat.reduce3.ent
- 1U72A.NA.fp.nowat.reduce3.ent
- placental.ancestral.2.1U70A.2.full2.reduce3.nohet.ent
- average.models.1.mmtp.full.reduce3.nowat.ent
- average.models.1.mmtp.full2.reduce3.nowat.ent
- average.models.1.nonrotamer.full.reduce3.nowat.ent
- average.models.1.nonrotamer.full2.reduce3.nowat.ent
- average.models.1.mmtp.full.reduce3.nohet.ent
- average.models.1.nonrotamer.full2.reduce3.nohet.ent
- mammal.chicken.AI.group1.full2.nowat.ent
- mammal.chicken.PI.GAO.group2.temp2.new3.nadph.vacuum2.reduce3.ent
- mammal.chicken.AG.GAO.new.full2.nohet.reduce3.ent
- mammal.chicken.AG.GAO.new.full.nohet.reduce3.ent
- mammal.chicken.AI.GAO.full2.nohet.reduce3.ent
- mammal.chicken.AI.GAO.full.nohet.reduce3.ent
- mammal.chicken.AIVY.full2.nohet.reduce3.ent
- mammal.chicken.AIVY.full.nohet.reduce3.ent
- mammal.chicken.PGVY.full2.nohet.reduce3.ent

[559] We would deposit these structures in the PDB, but it unfortunately no longer accepts non-experimental structure files.

- mammal.chicken.PGVY.full.nohet.reduce3.ent
- mammal.chicken.PGVYS.new.full2.nohet.reduce3.ent
- mammal.chicken.PGVYS.new.full.nohet.reduce3.ent
- mammal.chicken.PI.GAO.new.full2.nohet.reduce3.ent
- mammal.chicken.PI.GAO.new.full.nohet.reduce3.ent
- EED.NA.full.nowat.reduce3.ent
- EED.NA.full2.nowat.reduce3.ent
- EEK.NA.full.nowat.reduce3.ent
- EEK.NA.full2.nowat.reduce3.ent
- K_D.full.nowat.reduce3.ent
- K_D.full2.nowat.reduce3.ent
- K_K.CL.full.nowat.reduce3.ent
- K_K.CL.full2.nowat.reduce3.ent
- KED.NA.full.nowat.reduce3.ent
- KED.NA.full2.nowat.reduce3.ent
- KEK.CL.full.nowat.reduce3.ent
- KEK.CL.full2.nowat.reduce3.ent
- _KAGKFEDQ.full2.new.reduce3.nohet.ent
- _KAGKFEDQ.full.new.reduce3.nohet.ent
- _KAGKFEDQ.loop.13-27.full2.new.reduce3.nohet.ent
- _E_SKFEDQ.full2.new.reduce3.nohet.ent
- _E_SKFEDQ.full.new.reduce3.nohet.ent
- _E_SKFEDQ.loop.13-27.full.new.reduce3.nohet.ent
- _E_EDQ.NA.full.nowat.reduce3.ent
- _E_EDQ.NA.full2.nowat.reduce3.ent
- _E_GKFED.full.nowat.reduce3.ent
- _E_GKFED.full2.nowat.reduce3.ent

- _E_GKFEDQ.full.nowat.reduce3.ent
- _E_GKFEDQ.full2.nowat.reduce3.ent
- _E_SKFED.full.nowat.reduce3.ent
- _E_SKFED.full2.nowat.reduce3.ent
- _E_SKFEDQ.full.nowat.reduce3.ent
- _E_SKFEDQ.full2.nowat.reduce3.ent
- _E_SKFEDQ.loop.13-27.full.nowat.reduce3.ent
- _E_SKFEDQ.loop.13-27.full2.nowat.reduce3.ent
- _EAED.NA.full.nowat.reduce3.ent
- _EAED.NA.full2.nowat.reduce3.ent
- _EAEDQ.NA.full.nowat.reduce3.ent
- _EAEDQ.NA.full2.nowat.reduce3.ent
- _EAGKFED.full.nowat.reduce3.ent
- _EAGKFED.full2.nowat.reduce3.ent
- _EAGKFEDQ.full.nowat.reduce3.ent
- _EAGKFEDQ.full2.nowat.reduce3.ent
- _EASKFED.full.nowat.reduce3.ent
- _EASKFED.full2.nowat.reduce3.ent
- _EASKFEDQ.full.nowat.reduce3.ent
- _EASKFEDQ.full2.nowat.reduce3.ent
- _EE_D.NA.full.nowat.reduce3.ent
- _EE_D.NA.full2.nowat.reduce3.ent
- _EED.NA.full.nowat.reduce3.ent
- _EED.NA.full2.nowat.reduce3.ent
- _EEK.NA.full.nowat.reduce3.ent
- _EEK.NA.full2.nowat.reduce3.ent
- _K_D.full.nowat.reduce3.ent

- _K_D.full2.nowat.reduce3.ent
- _K_K.CL.full.nowat.reduce3.ent
- _K_K.CL.full2.nowat.reduce3.ent
- _KAED.NA.full.nowat.reduce3.ent
- _KAED.NA.full2.nowat.reduce3.ent
- _KAEDQ.NA.full.nowat.reduce3.ent
- _KAEDQ.NA.full2.nowat.reduce3.ent
- _KAGKFED.full.nowat.reduce3.ent
- _KAGKFED.full2.nowat.reduce3.ent
- _KAGKFEDQ.full.nowat.reduce3.ent
- _KAGKFEDQ.full2.nowat.reduce3.ent
- _KAGKFEDQ.loop.13-27.full.nowat.reduce3.ent
- _KAGKFEDQ.loop.13-27.full2.nowat.reduce3.ent
- _KASKFED.full.nowat.reduce3.ent
- _KASKFED.full2.nowat.reduce3.ent
- _KASKFEDQ.full.nowat.reduce3.ent
- _KASKFEDQ.full2.nowat.reduce3.ent
- _KAYGKFEDQ.f.full.nowat.reduce3.ent
- _KAYGKFEDQ.f.full2.nowat.reduce3.ent
- _KAYGKFEDQ.f2.full.nowat.reduce3.ent
- _KAYGKFEDQ.f2.full2.nowat.reduce3.ent
- _KAYGKFEDQ.loop.13-27.f2.full.nowat.reduce3.ent
- _KAYGKFEDQ.loop.13-27.f2.full2.nowat.reduce3.ent
- _KE_D.full.nowat.reduce3.ent
- _KE_D.full2.nowat.reduce3.ent
- _KED.NA.full.nowat.reduce3.ent
- _KED.NA.full2.nowat.reduce3.ent

- _KEK.CL.full.nowat.reduce3.ent
- _KEK.CL.full2.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.f2.f2p.fSA.fp.173529.nohet.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.f2.f2p.fSA.fp.173530.nohet.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.f2.f2p.fSA.fp.173531.nohet.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.fSA.full2.nohet.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.fSA.full.nohet.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.full2.nohet.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.full2.partial.nohet.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.full.nohet.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.full.partial.nohet.reduce3.ent
- fungi_metazoa.1111_GVF.S.na.fSA.full2.nohet.reduce3.ent
- fungi_metazoa.1111_GVF.S.na.fSA.full.nohet.reduce3.ent
- fungi_metazoa.1111_GVF.S.na.full2.nohet.reduce3.ent
- fungi_metazoa.1111_GVF.S.na.full.nohet.reduce3.ent
- fungi_metazoa.1111_STF.S.na.f2.f2p.fSA.fp.173529.nohet.reduce3.ent
- fungi_metazoa.1111_STF.S.na.f2.f2p.fSA.fp.173530.nohet.reduce3.ent
- fungi_metazoa.1111_STF.S.na.f2.f2p.fSA.fp.173531.nohet.reduce3.ent
- fungi_metazoa.1111_STF.S.na.fSA.full2.nohet.reduce3.ent
- fungi_metazoa.1111_STF.S.na.fSA.full.nohet.reduce3.ent
- fungi_metazoa.1111_STF.S.na.full2.nohet.reduce3.ent
- fungi_metazoa.1111_STF.S.na.full2.partial.nohet.reduce3.ent
- fungi_metazoa.1111_STF.S.na.full.nohet.reduce3.ent
- fungi_metazoa.1111_STF.S.na.full.partial.nohet.reduce3.ent
- fungi_metazoa.1111_SVF.S.na.fSA.full2.nohet.reduce3.ent
- fungi_metazoa.1111_SVF.S.na.fSA.full.nohet.reduce3.ent
- fungi_metazoa.1111_SVF.S.na.full2.nohet.reduce3.ent

- fungi_metazoa.1111_SVF.S.na.full.nohet.reduce3.ent
- fungi_metazoa.1111_chars2.idm.freeze1.new.reduce3.ent
- fungi_metazoa.1111_chars2.idm.freeze1.new.reduce3.ent
- fungi_metazoa.1111_chars2.mutated2.loop.2.nadph.new.reduce3.ent
- fungi_metazoa.1111_GTF.na.full.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.na.full2.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.f2.f2p.fSA.fp.173529.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.f2.f2p.fSA.fp.173530.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.f2.f2p.fSA.fp.173531.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.f2.fSA.f2p.173529.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.f2.fSA.f2p.173530.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.f2.fSA.f2p.173531.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.fpSA.f2p.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.fpSA.fp.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.fSA.full.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.fSA.full2.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.full.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.full.partial.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.full2.nowat.reduce3.ent
- fungi_metazoa.1111_GTF.S.na.full2.partial.nowat.reduce3.ent
- fungi_metazoa.1111_GVF.na.full.nowat.reduce3.ent
- fungi_metazoa.1111_GVF.na.full2.nowat.reduce3.ent
- fungi_metazoa.1111_GVF.S.na.fSA.full.nowat.reduce3.ent
- fungi_metazoa.1111_GVF.S.na.fSA.full2.nowat.reduce3.ent
- fungi_metazoa.1111_GVF.S.na.full.nowat.reduce3.ent
- fungi_metazoa.1111_GVF.S.na.full.partial.nowat.reduce3.ent
- fungi_metazoa.1111_GVF.S.na.full2.nowat.reduce3.ent

- fungi_metazoa.1111_GVF.S.na.full2.partial.nowat.reduce3.ent
- fungi_metazoa.1111_STF.na.full.nowat.reduce3.ent
- fungi_metazoa.1111_STF.na.full2.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.f2.f2p.fSA.fp.173529.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.f2.f2p.fSA.fp.173530.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.f2.f2p.fSA.fp.173531.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.f2.fSA.f2p.173529.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.f2.fSA.f2p.173530.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.f2.fSA.f2p.173531.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.fpSA.f2p.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.fpSA.fp.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.fSA.full.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.fSA.full2.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.full.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.full.partial.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.full2.nowat.reduce3.ent
- fungi_metazoa.1111_STF.S.na.full2.partial.nowat.reduce3.ent
- fungi_metazoa.1111_SVF.na.full.nowat.reduce3.ent
- fungi_metazoa.1111_SVF.na.full2.nowat.reduce3.ent
- fungi_metazoa.1111_SVF.S.na.fSA.full.nowat.reduce3.ent
- fungi_metazoa.1111_SVF.S.na.fSA.full2.nowat.reduce3.ent
- fungi_metazoa.1111_SVF.S.na.full.nowat.reduce3.ent
- fungi_metazoa.1111_SVF.S.na.full.partial.nowat.reduce3.ent
- fungi_metazoa.1111_SVF.S.na.full2.nowat.reduce3.ent
- fungi_metazoa.1111_SVF.S.na.full2.partial.nowat.reduce3.ent
- ascomycota.0010_KD.vacuum.new.reduce3.ent
- ascomycota.0010_KD.vacuum.new.reduce3.ent

- ascomycota.0010_KP.vacuum2.new.reduce3.ent
- ascomycota.0010_KP.vacuum2.new.reduce3.ent
- ascomycota.0010_QD.vacuum2.new.reduce3.ent
- ascomycota.0010_QD.vacuum2.new.reduce3.ent
- ascomycota.0010_QP.vacuum2.new.reduce3.ent
- ascomycota.0010_QP.vacuum2.new.reduce3.ent
- ascomycota.0011_KD.na.fp.nowat.reduce3.ent
- ascomycota.0011_KD.na.full.nowat.reduce3.ent
- ascomycota.0011_KD.na.full2.nowat.reduce3.ent
- ascomycota.0011_KD.vacuum.new.reduce3.ent
- ascomycota.0011_KD.vacuum2.new.reduce3.ent
- ascomycota.0011_KP.fp.nowat.reduce3.ent
- ascomycota.0011_KP.full.nowat.reduce3.ent
- ascomycota.0011_KP.full2.nowat.reduce3.ent
- ascomycota.0011_QD.na.fp.nowat.reduce3.ent
- ascomycota.0011_QD.na.full.nowat.reduce3.ent
- ascomycota.0011_QD.na.full2.nowat.reduce3.ent
- ascomycota.0011_QD.vacuum.new.reduce3.ent
- ascomycota.0011_QD.vacuum2.new.reduce3.ent
- ascomycota.0011_QP.na.fp.nowat.reduce3.ent
- ascomycota.0011_QP.na.full.nowat.reduce3.ent
- ascomycota.0011_QP.na.full2.nowat.reduce3.ent

Appendix N: COPYING

The GNU Affero GPL, version 3 (Foundation 2007) license (or a later version of it, at your option) covers all programs in, including as supplemental files of or (if written by the author of the dissertation) mentioned in, this dissertation. It is available online via http://www.gnu.org/licenses/ and at http://cesario.rutgers.edu/easmith/research/perl/COPYING.

Material other than programs in this dissertation, including (if created by the author of this dissertation) supplemental files and material available online via URLs (e.g., http://cesario.rutgers.edu/easmith/research/) mentioned in this dissertation, is available under a Creative Commons Attribution-ShareAlike 2.5 License (Commons, C 2006). It is available online via http://creativecommons.org/licenses/by-sa/2.5/ and, in its “legal code” version, at http://cesario.rutgers.edu/easmith/research/legalcode.txt.

Appendix O: Outgroup review/explanation

Outgroups are species, or groups of species, that are known[560] to be outside of[561] the group of species of interest, but are used to help in phylogenetic work in two major areas:

- In determining ancestral sequences; without an outgroup sequence or sequences, one would have difficulty determining what the ancestral sequence was in locations in which the present-day sequences are variable (Edwards, R J & Shields 2004)[562]. For example, if part of a (fictional) protein was "ICQW" in mouse and "ISQF" in human, then it would be difficult to predict whether the second residue was "S" or "C" (and similarly for the last residue. If one also had the sequence from, e.g., chicken (the outgroup), and it was "VCNF", then the likeliest sequence in the ancestor of mice and men would be ICQF. Please see Figure O.1, on page 413, for a graphical representation of this.

[560] Evidence that the outgroup species are not inside the group of species of interest should ideally (Simmons *et al.* 2002) be available from sources external to the sequences currently under study (e.g., from fossil evidence). Mistakes in this determination can be problematic (Robinson, M *et al.* 1998; Van de Peer *et al.* 2002).

[561] By "outside of" is meant "branching off prior to all of the species of interest".

[562] For one example from the present work, note that it would have been difficult to determine, for instance, the Archaea/Eukaryota ancestral sequence for ORO (see "Central protein candidates", on page 48) without using species from all of Archaea, Bacteria, and Eukaryota. This concern was one reason that ORO was not selected.

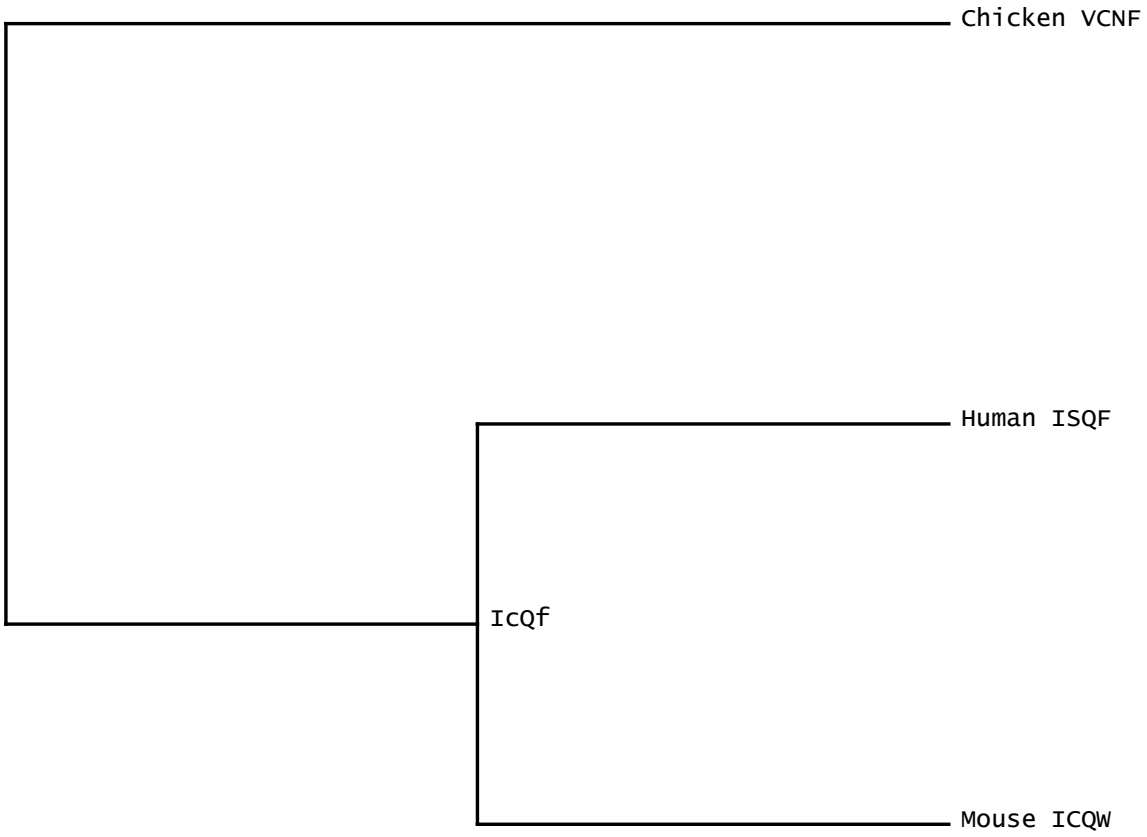


Figure O.1: Outgroup sequence example

- Outgroups also have other uses (Cotter, Caffrey, & Shields 2002; Fuellen, Wagele, & Giegerich 2001; Glazko & Nei 2003), including:
  - For "rooting" a tree (Huelsenbeck, Bollback, & Levine 2002), as in detecting what branch is the most ancestral one[563] (without making an assumption of a molecular clock);
  - Their use in the *implicit* determination of ancestral sequences, since methods of tree creation (other than those purely based on pair distances) determine ancestral sequences as a part of their algorithms.

[563] One use for this is to determine where the tree can be best attached to other trees, as with the tree assembly process - see "3a. Creation of a rough starting tree", on page 72.

Outgroups, however, do have some potential problems associated with them, particularly if they are too distant from the species of interest (Dacks *et al.* 2002; Graham, Olmstead, & Barrett 2002; Lartillot, Brinkmann, & Philippe 2007; Moreira, Lopez-Garcia, & Vickerman 2004; Philippe, Lartillot, & Brinkmann 2005; Simmons *et al.* 2002; Tarrio, Rodriguez-Trelles, & Ayala 2000). Problems with overly distant outgroups range from alignment difficulties to long branch attraction effects (see footnote 52 under "Tree construction methods", on page 27). Depending on whether the outgroup is constrained to be at the root of the tree, long branch attraction to an outgroup tends to cause either:

- The outgroup moving upward into the tree and long branched non-outgroup species moving downward toward the outgroup's (incorrect) position, if the outgroup is not constrained. This problem appears to have happened in the present research with some of the tree searches, such as "Tree search with Eukaryota (subset)", on page 300, for the Archaea (composite outgroup sequence) and *Tetrahymena thermophila*.
- Species with long branches moving downward if the outgroup is constrained, thus appearing more basal (i.e., branching off earlier) than they actually are (Dacks *et al.* 2002).

Appendix P: Perl programs created

The below is a listing of the 180 Perl (Wall, Christiansen, & Orwant 2000) programs created for this dissertation (not including those superseded by the below). They are available in the supplemental file "perl.tar", which is an archive in UNIX tar format, and under http://cesario.rutgers.edu/easmith/research/perl/. All are licensed under the AGPL (see "Appendix N: COPYING", on page 411). Note that some programs are not directly mentioned in the text, but are used in intermediary or data-extraction steps via the "`make`" (Stallman, McGrath, & Smith 1998) program[564], are run by other programs, or are for display purposes. A few programs have the same source code but differ in how they run depending on the name by which they are called; this is denoted by the names being together in a list with an "or" (and is implemented in UNIX/Linux operating systems by the creation of "symlinks"). They are in alphabetical order.

- 3d_ali.extract.aligned2.pl
- 3d_ali.extract.pl
- add.restraints.wrapper.pl
- align.all.seqs.to.chains.pl
- align.atom.pdb.seqs.pl
- align.caccts.1.pl
- align.caccts.2.pl
- align.chain.to.seqs.pl
- align.clustered.pdbs.2.pl

[564] This program uses either "Makefile.prior" (for earlier stages) or "Makefile" (for later stages); these are available as supplementary files with a ".txt" ending and in the directory noted above.

- align.clustered.pdbs.3.pl
- align.clustered.pdbs.pl
- align.consensus.chain.to.seqs.pl
- align.lsqrms.wrapper.full.pl or align.lsqrms.wrapper.pl
- align.nr.chain.to.seqs.pl
- align.nr.partial.chain.to.seqs.pl
- align.polymorphism.add.seqs.pl
- align.polymorphism.pl
- align.to.central.2.pl
- align.to.central.3.pl
- analyze.align.clustered.pdbs.10.pl
- analyze.align.clustered.pdbs.11.pl
- analyze.align.clustered.pdbs.2.pl
- analyze.align.clustered.pdbs.3.pl
- analyze.align.clustered.pdbs.4.pl
- analyze.align.clustered.pdbs.5.pl
- analyze.align.clustered.pdbs.6.pl
- analyze.align.clustered.pdbs.9.pl
- analyze.align.to.central.1.pl
- analyze.align.to.central.2.pl
- average.hetatm.4.pl
- caccts.2.pl
- caccts.3.pl
- check.atom.distances4.pl
- check.atom.distances5.pl
- check.for.bumps.2.pl
- check.main.chain.distances.2.pl

- check.pdb.vs.pfam.pl
- check.side.chain.distances.1.pl
- check.single.structural.align.pl
- check.water.for.pos.ions.pl
- cluster.chain.to.seq.pl
- combine.align.to.central.3.pl
- combine.structural.align.groups.pl
- compare.trees.problems.pl
- consensus.2.chain.to.seqs.pl
- consensus.2.nr.chain.to.seqs.pl
- consensus.align.consensus.chain.to.seqs.pl
- consensus.chain.to.seqs.pl
- consensus.nr.chain.to.seqs.pl
- consensus4.multiple.pl
- consensus5.multiple.pl
- convert.structal.pl
- convert.structal.pl
- create.freezegrps.2.pl
- create.ins.del.mutate.freezegrps.2.pl
- create.mkrotscr.mutate.1.loop.pl
- create.mkrotscr.mutate.2.loop.pl
- create.outgroup.seqs.pl
- create.restraints.1.pl
- create.restraints.2.pl
- create.similarity.matrix.1.pl
- create.tree.section.pl
- determine.parsimony.species.pl

- do.reduce3.restore.pl
- estimate.starting.dists.3.pl
- extract.30.pairs.pl
- extract.atom.seqs.pl
- extract.chain.records.pl
- extract.important.pdbs.pl
- extract.needed.pdbs.for.loop.search.pl
- extract.protein.clusters.pl
- extract.r.values.pl
- extract.species.names.pl
- extract.sptrembl.polymorphism.pl
- figure.out.kingdom.norm.dists.pl
- find.align.thresholds.pl
- find.AO.NO.AOP.constraints.2.pl
- find.distance.deviations.2.pl
- find.distance.min.max.matrices.pl
- find.helix.coil.sub.groups.pl
- find.interacting.res.pl
- find.pdbatom.stockholm.alignment.pl
- find.residue.correlations.2.pl
- find.residue.correlations.pl
- find.species.weights.2.pl
- find.species.weights.3.pl
- find.sub.groups.pl
- group.to.msf.pl
- group.to.msf.spread.pl
- homstrad.extract.pl

- integrate.sequence.align.1.pl
- integrate.structural.align.1.pl
- integrate.structural.align.2.pl
- interpret.align.pdbs.pl
- interpret.important.pdbs.pl
- interpret.pdb.clusters.pl
- interpret.probe.pl
- interpret.protein.files.pl
- interpret.ring.changes.pl
- list.polymorphism.pl
- nexus.add.freqs.pl
- nexus.add.gap.partitions.pl
- nexus.add.groups.2.pl
- nexus.add.kingdom.constraints.pl
- nexus.add.usertree.section.pl
- nexus.add.wag.pl
- nexus.cleanup.quartets.pl
- nexus.consolidate.partitions.pl
- nexus.create.ncbi.subtrees.10.pl
- nexus.create.ncbi.subtrees.2.pl
- nexus.create.ncbi.subtrees.3.pl
- nexus.create.ncbi.subtrees.4.pl
- nexus.create.ncbi.subtrees.5.pl
- nexus.create.ncbi.subtrees.6.pl
- nexus.create.ncbi.subtrees.7.pl
- nexus.create.ncbi.subtrees.8.pl
- nexus.create.ncbi.subtrees.9.pl

- nexus.create.ncbi.subtrees.MyTree0001.pl
- nexus.create.ncbi.subtrees.pl
- nexus.create.ncbi.tree.pl
- nexus.extract.ancestral.seq.pos.pl
- nexus.extract.ancestral.seqs.2.pl
- nexus.extract.ancestral.seqs.3.pl
- nexus.find.init.quartets.pl
- nexus.find.overall.quartets.1.pl
- nexus.get.quartets.2.pl
- nexus.get.quartets.2.wrapper.pl
- nexus.get.quartets.kingdom.pl
- nexus.get.quartets.pl
- nexus.get.quartets.recover.pl
- nexus.get.quartets.recover2.pl
- nexus.get.quartets.recover3.pl
- nexus.interpret.TreeBASE.trees.pl
- nexus.make.charset.seqs.pl
- nexus.remove.excluded.pl
- nexus.simplify.full.species.2.pl
- nexus.simplify.full.species.pl
- nexus.simplify.polymorphism.pl or nexus.simplify.polymorphism.all.pl
- nexus.split.non_metazoa.quartets.pl
- nexus.use.recdcm3.subsets.pl
- pdb.get.chains.pl
- pdb.nr.embl.swissprot.pl
- pdb.nr.multiple.pl
- pdb.nr.multiple2.pl

- pdb.replaced.groups.pl
- pdb.species.extract.pl
- pdbeast.species.extract2.pl
- pfam.swissprot.to.pdb.pl
- process.align.to.central.pl
- process.structural.align.pl
- process2.structural.align.pl
- protein.replaced.groups2.pl
- put.dists.on.tree.pl
- put.together.pdbs.section.3.pl
- put.together.pdbs.section.4.pl
- put.together.pdbs.sequence.3.pl
- quartets.to.weights.pl
- quartets.to.wr.modified.pl
- recdcm3.get.subsets.pl
- reformat.gromacs.pdb.pl
- reformat.pdb.gromacs.pl
- restore.reduce3.removed.pl
- ring.changes.lsqrms.pl
- rotrans.lsqrms.pl
- run.lsqrms.pl, run.lsqrms.simple.pl, or run.lsqrms.simple.full.pl
- select.aligned.important.pdbs.pl
- split.showalign.html.pl
- sprot.group.protein.files.pl
- sprot.pdb.DR.extract2.pl
- sprot.pdb.species.extract.pl
- summarize.chain.lengths.pl

- sump.summarize.pl
- sump.summarize2.pl
- swissprot.scop.species2.pl
- swissprot.species3.filter3.pl
- test.find.quartets.1.pl
- to.gromacs.wrapper.2.pl
- transfer.weights.to.stockholm.1.pl
- tree.simplify.full.pl
- use.mrbayes.sump.freqs.info.2.pl
- use.mrbayes.sump.freqs.info.pl

Appendix Q: Non-local programs used/mentioned

The programs (including groups of programs) listed in this appendix are all of non-local creation, except that some (as noted) have been modified locally. Those modified in significant aspects (i.e., which made a difference in the research), except for purely as necessary to get them to compile and run on the local machines:

- GROMACS (Berendsen, van der Spoel, & van Drunen 1995; Lindahl, Hess, & van der Spoel 2001; Lindahl *et al.* 2007; van der Spoel *et al.* 2005); the modifications were to:
  - The "genbox" program, via a patch of "addconf.c" (http://cesario.rutgers.edu/easmith/research/patches/addconf.c.patch)
  - Some datafiles (in the "top" directory):
    - http://cesario.rutgers.edu/easmith/research/patches/ffG43b1.hdb.patch
    - http://cesario.rutgers.edu/easmith/research/patches/ffG53a6.hdb.patch
    - http://cesario.rutgers.edu/easmith/research/patches/vdwradii.dat.patch
    - http://cesario.rutgers.edu/easmith/research/patches/xlateat.dat.patch
- HMMer (Eddy & Birney 2003); the modifications were to the "hmmbuild" program - see http://cesario.rutgers.edu/easmith/research/patches/hmmbuild.c.patch
- LSQRMS (Alexandrov & Graham 2003; Gerstein & Levitt 1996, 1998); see http://cesario.rutgers.edu/easmith/research/patches/lsqrms-2.0.4b.patch
- MrBayes (Altekar *et al.* 2004; Huelsenbeck & Ronquist 2001; Huelsenbeck *et al.* 2006; Ronquist & Huelsenbeck 2003); see http://cesario.rutgers.edu/easmith/research/patches/mrbayes-3.1.2.patch

Those not modified in significant aspects (i.e., except as necessary for functioning on the local machines, if applicable):

- blastp (Altschul *et al.* 1990; Altschul *et al.* 1997; Gertz 2006)

- ClustalW (Thompson, J D, Higgins, & Gibson 1994)
- dang (Word 2000)
- GROMACS:
  - editconf
  - g_disre
  - grompp
  - mdrun
  - pdb2gmx
- HMMer:
  - hmmalign
  - hmmemit
  - sreformat
- KiNG (Richardson, D C 2007)
- `make` (Stallman, McGrath, & Smith 1998)
- MolProbity (Davis *et al.* 2007; Lovell *et al.* 2003)
- PHYLIP (Felsenstein 1993):
  - CONSENSE
  - FITCH
  - PENNY
- prekin, `probe`, and mkrotscr (Word *et al.* 2000)
- QuartetSuite (Piaggio-Talice & Piaggio 2003; Piaggio-Talice, Burleigh, & Eulenstein 2004):
  - Assemble
  - Rectify
- `reduce` (Word *et al.* 1999a; Word *et al.* 1999b; Word & Richardson 2006)
- scanprosite (de Castro *et al.* 2006)
- Tree-Puzzle (von Haeseler & Strimmer 2003; Schmidt *et al.* 2002; Strimmer & von Haeseler 1996, 1999)

The following programs were mentioned in the text, but not used:

- ACCESS (Lee, B-K & Richards 1971)
- AtVol (Word 1999)
- calc-volume (Gerstein & Richards 2001; Tsai *et al.* 1999)
- GROMACS: genion
- Modeller (Fiser, Do, & Sali 2000; Sali & Blundell 1993; Sali & Overington 1994; Sali 1995, 2001)
- SWISS-MODEL (Schwede *et al.* 2003)
- VOLUME (Biology 2006; Richards 1974)

Appendix R: Supplemental files and URLs

Listed below are the supplemental files and corresponding URLs for this dissertation. Any file ending with ".tar" is a UNIX "tar" archive containing multiple files.

- Alignments of non-DHFR/TS proteins (see "3b. Alignment of other sequences" on page 194):
  - http://cesario.rutgers.edu/easmith/research/alignments/
  - Supplemental file: alignments.tar
- NADPH/DHFR constraints file (see "Creation of restraints" on page 170):
  - http://cesario.rutgers.edu/easmith/research/AO.NO.constraints.ascomycota.txt
  - Supplemental file: AO.NO.constraints.ascomycota.txt
- Partial DHFR alignment (see "Appendix K: Partial DHFR alignment" on page 384):
  - http://cesario.rutgers.edu/easmith/research/DHFR.with.fungi.2.seqs.edited.7.vert.xls
  - Supplemental file: DHFR.with.fungi.2.seqs.edited.7.vert.xls
- Full DHFR alignment (see "5. Alignment of central sequences" on page 128 and on page 336):
  - http://cesario.rutgers.edu/easmith/research/DHFR.with.fungi.2.stockholm.txt
  - Supplemental file: DHFR.with.fungi.2.stockholm.txt
- GROMACS' ".mdp" files (see "7. Model building" on page 146):
  - http://cesario.rutgers.edu/easmith/research/mdp/
  - Supplemental file: mdp.tar
- MolProbity output files (see "Appendix E: MolProbity results" on page 371):
  - Summary:
    - http://cesario.rutgers.edu/easmith/research/molprobity/extract.molprobity.1.new.xls
    - Supplemental file: extract.molprobity.1.new.xls
  - Files:
    - http://cesario.rutgers.edu/easmith/research/molprobity/
    - Supplemental file: molprobity.html.tar

- Perl (see “Appendix P: Perl programs created” on page 415):
  - First makefile, used by `make` (Stallman, McGrath, & Smith 1998):
    - http://cesario.rutgers.edu/easmith/research/perl/Makefile.prior
    - Supplemental file: Makefile.prior.txt
  - Second makefile:
    - http://cesario.rutgers.edu/easmith/research/perl/Makefile
    - Supplemental file: Makefile.txt
  - All perl programs:
    - http://cesario.rutgers.edu/easmith/research/perl/
    - Supplemental file: perl.tar
- Proteins used:
  - PDB files examined/used (see “Appendix A: PDB files/chains used” on page 366 and “Appendix B: Important PDB files/chains used” on page 367):
    - http://cesario.rutgers.edu/easmith/research/proteins/extract.important.pdbs.txt
    - Supplemental file: extract.important.pdbs.txt
    - http://cesario.rutgers.edu/easmith/research/proteins/extract.important.pdbs.xls
    - Supplemental file: extract.important.pdbs.xls
    - http://cesario.rutgers.edu/easmith/research/proteins/interpret.important.pdbs.xls
    - Supplemental file: interpret.important.pdbs.xls
    - http://cesario.rutgers.edu/easmith/research/proteins/interpret.important.pdbs.txt.new
    - Supplemental file: interpret.important.pdbs.txt.new.txt
  - Sequences and structures used (see “Selection of structures and other sequences” on page 51 and “Structures and sequences” on page 61):
    - http://cesario.rutgers.edu/easmith/research/proteins/important.protein.files.all.txt.new
    - Supplemental file: important.protein.files.all.txt.new.txt
    - http://cesario.rutgers.edu/easmith/research/proteins/interpret.protein.files.txt.new
    - Supplemental file: interpret.protein.files.txt.new.txt

- Polymorphism (see “Criteria for polymorphic sequences used” on page 65):
  - http://cesario.rutgers.edu/easmith/research/proteins/extract.sptrembl.polymorphism.txt
  - Supplemental file: extract.sptrembl.polymorphism.txt
  - http://cesario.rutgers.edu/easmith/research/proteins/list.polymorphism.txt
  - Supplemental file: list.polymorphism.txt
  - http://cesario.rutgers.edu/easmith/research/proteins/proteins.polymorphism.manual.txt
  - Supplemental file: proteins.polymorphism.manual.txt
  - http://cesario.rutgers.edu/easmith/research/proteins/split.polymorphism.txt
  - Supplemental file: split.polymorphism.txt
- Input files for put.together.pdbs.* programs (see “Assignment of initial coordinates” on page 150 and “Loop searches” on page 157):
  - http://cesario.rutgers.edu/easmith/research/put.together.pdbs/
  - Supplemental file: put.together.pdbs.tar
- Species data:
  - Species ambiguities (see “Resolution of species ambiguities” on page 77 and “Appendix D: NCBI taxids and alternate species names” on page 370):
    - Summary:
      - http://cesario.rutgers.edu/easmith/research/species/extract.species.used.taxdump.data.xls
      - Supplemental file: extract.species.used.taxdump.data.xls
    - Full database:
      - http://cesario.rutgers.edu/easmith/research/species/bad.nodes.txt
      - Supplemental file: bad.nodes.txt
      - http://cesario.rutgers.edu/easmith/research/species/genus.above.names.NCBI.txt
      - Supplemental file: genus.above.names.NCBI.txt
      - http://cesario.rutgers.edu/easmith/research/species/species.lineage.NCBI.txt
      - Supplemental file: species.lineage.NCBI.txt
      - http://cesario.rutgers.edu/easmith/research/species/species.names.NCBI.txt
      - Supplemental file: species.names.NCBI.txt

        - http://cesario.rutgers.edu/easmith/research/species/species.subspecies.NCBI.txt
        - Supplemental file: species.subspecies.NCBI.txt
    - Species versus structures (see “Database of structures and species” on page 55 and “2. Determine sources for phylogenetic sequences” on page 191):
        - http://cesario.rutgers.edu/easmith/research/species/known.species.txt
        - Supplemental file: known.species.txt
        - http://cesario.rutgers.edu/easmith/research/species/swissprot.scop.species2.txt
        - Supplemental file: swissprot.scop.species2.txt
    - Groups of species (see “Appendix I: Species groupings used” on page 376):
        - http://cesario.rutgers.edu/easmith/research/species/species.groups.txt
        - Supplemental file: species.groups.txt
- Model structures (see “Appendix M: Model PDB-format files” on page 403):
    - http://cesario.rutgers.edu/easmith/research/struct/
    - Supplemental file: struct.tar
- Trees:
    - “Parsimony” tree (see “Initial sources” on page 72):
        - http://cesario.rutgers.edu/easmith/research/trees/MyTree0001.nexus
        - Supplemental file: MyTree0001.nexus.txt
        - http://cesario.rutgers.edu/easmith/research/trees/weights.single.txt
        - Supplemental file: weights.single.txt
    - TreeBASE trees used (see “Usage of quartets” on page 74):
        - http://cesario.rutgers.edu/easmith/research/trees/TreeBASE.trees.used.txt
        - Supplemental file: TreeBASE.trees.used.txt
    - Tree results (see “Appendix L: Tree files available, cross-referenced to pictures” on page 394)
        - http://cesario.rutgers.edu/easmith/research/trees/
        - Supplemental file: trees.tar

- Simulated Annealing (SA) results (see “Simulated Annealing (SA)” on page 195):
  - http://cesario.rutgers.edu/easmith/research/new.SA.archaea.xls
  - Supplemental file: new.SA.archaea.xls
  - http://cesario.rutgers.edu/easmith/research/trees/new.SA.bacteria.xls
  - Supplemental file: new.SA.bacteria.xls
  - http://cesario.rutgers.edu/easmith/research/trees/new.SA.eukaryota.xls
  - Supplemental file: new.SA.eukaryota.xls

Works Cited

Please note that the citations are sorted by the first author (without any “de” or other uncapitalized portions) then the year.

Aagaard, C; Douthwaite, S. 1994 Apr 12. Requirement for a conserved tertiary interaction in the core of 23S ribosomal RNA. *Proceedings of the National Academy of Sciences USA* 91(8):2989-2993. http://www.pnas.org/cgi/content/abstract/91/8/2989.

Abagyan, R A; Batalov, S. 1997 Oct 17. Do aligned sequences share the same fold? *Journal of Molecular Biology* 273(1):355-368.

Aberqel, C; Coutard, B; Byrne, D; Chenivesse, S; Claude, J B; Dereqnaucourt, C; Fricaux, T; Gianesini-Boutreux, C; Jeudy, S; Lebrun, R; Maza, C; Notredame, C; Poirot, O; Suhre, K; Varaqnol, M; Claverie, J-M. 2003 Jun. Structural genomics of highly conserved microbial genes of unknown function in search of new antibacterial targets. *Journal of Structural and Functional Genomics* 4(2-3):141-157.

Acheson, S A; Bell, J B; Jones, M E; Wolfenden, R. 1990 Apr 3. Orotidine-5'-monophosphate decarboxylase catalysis: Kinetic isotope effects and the state of hybridization of a bound transition-state analogue. *Biochemistry* 29(13):3198-3202.

Adams, T. 1614. Change, n. "*The divells banket described in sixe sermons*" cited in *Oxford English Dictionary* ed: Simpson, J. Edition: 2nd (1989). Oxford University Press (Oxford).

Adey, N B; Tollefsbol, T O; Sparks, A B; Edgell, M H; Huchinson, C A, III. 1994 Feb 15. Molecular resurrection of an extinct ancestral promoter for mouse L1. *Proceedings of the National Academy of Sciences USA* 91(4):1569-1573. http://www.pnas.org/cgi/content/abstract/91/4/1569.

Alexandrov, V; Graham, J. 2003. LSQRMS, 2.04b. Yale University (New Haven, Connecticut). http://geometry.molmovdb.org/3dhmm/.

Alroy, J. 1995 Jun. Continuous track analysis: A new phylogenetic and biogeographic method. *Systematic Biology* 44(2):152-178.

Altekar, G; Dwarkadas, S; Huelsenbeck, J P; Ronquist, F. 2004 Feb 12. Parallel Metropolis-coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20(3):407-415. http://bioinformatics.oxfordjournals.org/cgi/content/abstract/20/3/407.

Altschul, S F; Carroll, R J; Lipman, D J. 1989 Jun 20. Weights for data related by a tree. *Journal of Molecular Biology* 207(4):647-653.

Altschul, S F; Gish, W; Miller, W; Myers, E W; Lipman, D J. 1990 Oct 5. Basic local alignment search tool. *Journal of Molecular Biology* 215(3):403-410.

Altschul, S F; Madden, T L; Schaffer, A A; Zhang, J; Zhang, Z; Miller, W; Lipman, D J. 1997 Sep 1. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 25(17):3389-3402. http://nar.oxfordjournals.org/cgi/content/full/25/17/3389.

Andersen, C A F; Palmer, A G; Brunak, S; Rost, B. 2002 Feb. Continuum secondary structure captures protein flexibility. *Structure (Cambridge)* 10(2):175-184.

Anderson, F E; Swofford, D L. 2004 Nov. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. *Molecular Phylogenetics and Evolution* 33(2):440-451.

Angen, O; Ahrens, P; Kuhnert, P; Christensen, H; Mutters, R. 2003 Sep 1. Proposal of *Histophilus somni* gen. nov., sp. nov. for the three species incertae sedis '*Haemophilus somnus*', '*Haemophilus agni*' and '*Histophilus ovis*'. *International Journal of Systematic and Evolutionary Microbiology* 53(5):1449-1456. http://ijs.sgmjournals.org/cgi/content/abstract/53/5/1449.

Appleby, T C; Kinsland, C; Begley, T P; Ealick, S E. 2000 Feb 29. The crystal structure and mechanism of orotidine 5'-monophosphate decarboxylase. *Proceedings of the National Academy of Sciences USA* 97(5):2005-2010. http://www.pnas.org/cgi/content/full/97/5/2005.

Appleman, J R; Howell, E E; Kraut, J; Kuhl, M; Blakley, R L. 1988a Jul 5. Role of aspartate 27 in the binding of methotrexate to dihydrofolate reductase from *Escherichia coli*. *Journal of Biological Chemistry* 263(19):9187-9198. http://www.jbc.org/cgi/content/abstract/263/19/9187.

Appleman, J R; Prendergast, N J; Delcamp, T J; Freisheim, J H; Blakley, R L. 1988b Jul 25. Kinetics of the formation and isomerization of methotrexate complexes of recombinant human dihydrofolate reductase. *Journal of Biological Chemistry* 263(21):10304-10313. http://www.jbc.org/cgi/content/abstract/263/21/10304.

Appleman, J R; Howell, E E; Kraut, J; Blakley, R L. 1990 Apr 5. Role of aspartate 27 of dihydrofolate reductase from *Escherichia coli* in interconversion of active and inactive enzyme conformers and binding of NADPH. *Journal of Biological Chemistry* 265(10):5579-5584. http://www.jbc.org/cgi/content/abstract/265/10/5579.

Ardley, H C; Moynihan, T P; Markham, A F; Robinson, P A. 2000 Apr 25. Promoter analysis of the human ubiquitin-conjugating enzyme gene family UBE2L1–4, including UBE2L3 which encodes UbcH7. *Biochimica et Biophysica Acta: Gene Structure and Expression* 1491(1-3):57-64.

Arvestad, L. 1997. Aligning coding DNA in the presence of frame-shifts error. pp 180-190 in *Combinatorial Pattern Matching*. Lecture Notes in Computer Science. Vol: 1264. http://citeseer.ist.psu.edu/arvestad97aligning.html http://www.nada.kth.se/~arve/cpm97.pdf.

Arvestad, L. 1999. *Algorithms for biological sequence alignment*. Dissertation. KTH Royal Institute of Technology (Stockholm, Sweden): Numerical Analysis and Computer Science. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-2905 http://www.nada.kth.se/~arve/publications/.

Arvestad, L; Berglund, A-C; Lagergren, J; Sennblad, B. 2003 Jul 3. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 19(Suppl 1):i7-i15. http://www.bioinformatics.oupjournals.org/cgi/content/abstract/19/suppl_1/i7.

Asara, J M; Schweitzer, M H; Freimark, L M; Phillips, M; Cantley, L C. 2007 13 Apr. Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry. *Science* 316(5822):280-285.

Aszodi, A; Munro, R E J; Taylor, W R. 1997. Protein modeling by multiple sequence threading and distance geometry. *PROTEINS: Structure, Function, and Genetics* 29(Suppl 1):38-42.

Azarya-Sprinzak, E; Naor, D; Wolfson, H J; Nussinov, R. 1997 Oct 1. Interchanges of spatially neighbouring residues in structurally conserved environments. *Protein Engineering* 10(10):1109-1122.

Baccanari, D P; Tansilk, R L; Joyner, S S; Fling, M E; Smith, P L; Freisheim, J H. 1989 Jan. Characterization of *Candida albicans* dihydrofolate reductase. *Journal of Biological Chemistry* 264(2):1100-1107. http://www.jbc.org/cgi/content/abstract/264/2/1100.

de Bakker, P I W; Bateman, A; Burke, D F; Miguel, R N; Mizuguchi, K; Si, J S; Shirai, H; Blundell, T L. 2001 Aug. HOMSTRAD: Adding sequence information to structure-based alignments of homologous protein families. *Bioinformatics* 17(8):748-749. http://www-cryst.bioc.cam.ac.uk/homstrad/ http://bioinformatics.oupjournals.org/cgi/content/abstract/17/8/748.

Baldauf, S L; Palmer, J D. 1993 Dec 15. Animals and fungi are each other's closest relatives: Congruent evidence from multiple proteins. *Proceedings of the National Academy of Sciences USA* 90(24):11558-11562. http://www.pnas.org/cgi/content/abstract/90/24/11558.

Baldauf, S L; Doolittle, W F. 1997 Oct 28. Origin and evolution of the slime molds (Mycetozoa). *Proceedings of the National Academy of Sciences USA* 94(22):12007-12012. http://www.pnas.org/cgi/content/abstract/94/22/12007.

Bapteste, E; Brinkmann, H; Lee, J A; Moore, D V; Sensen, C W; Gordon, P; Durufle, L; Gaasterland, T; Lopez, P; Muller, M; Philippe, H. 2002 Feb 5. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proceedings of the National Academy of Sciences USA* 99(3):1414-1419. http://www.pnas.org/cgi/content/full/99/3/1414.

Bateman, A; Birney, E; Cerruti, L; Durbin, R M; Etwiller, L; Eddy, S R; Griffiths-Jones, S R; Howe, K L; Marshall, M; Sonnhammer, E L L. 2002 Jan 1. The Pfam Protein Families database. *Nucleic Acids Research* 30(1):276-280. http://nar.oupjournals.org/cgi/content/full/30/1/276 http://pfam.janelia.org/.

Begley, T P; Appleby, T C; Ealick, S E. 2000 Dec 1. The structural basis for the remarkable catalytic proficiency of orotidine 5'-monophosphate decarboxylase. *Current Opinion in Structural Biology* 10(6):711-718.

Bell, J B; Jones, M E. 1991 Jul 5. Purification and characterization of yeast orotidine 5'-monophosphate decarboxylase overexpressed from plasmid PGU2. *Journal of Biological Chemistry* 266(19):12662-12667. http://www.jbc.org/cgi/content/abstract/266/19/12662.

Benner, S A; Cannarozzi, G; Gerloff, D L; Turcotte, M; Chelvanayagam, G. 1997 Dec. Bona fide predictions of protein secondary structure using transparent analyses of multiple sequence alignments. *Chemical Reviews* 97(8):2725-2843.

Bensasson, D; Zhang, D-X; Hewitt, G M. 2000 Mar. Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. *Molecular Biology and Evolution* 17(3):406-415. http://mbe.oupjournals.org/cgi/content/full/17/3/406.

Benson, D A; Karsch-Mizrachi, I; Lipman, D J; Ostell, J; Rapp, B A; Wheeler, D L. 2000 Jan 1. GenBank. *Nucleic Acids Research* 28(1):15-18. http://nar.oxfordjournals.org/cgi/content/full/28/1/15.

Berendsen, H J C; Postma, J P M; van Gunsteren, W F; Hermans, J. 1981. Interaction models for water in relation to protein hydration. pp 331-342 in *Intermolecular Forces* ed: Pullman, B. D. Reidel Publishing Company (Dordrecht, Netherlands).

Berendsen, H J C; Postma, J P M; van Gunsteren, W F; DiNola, A; Haak, J R. 1984 Oct 15. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics* 81(8):3684-3690.

Berendsen, H J C; van der Spoel, D; van Drunen, R. 1995 Sep. GROMACS: A message-passing parallel molecular dynamics implementation. *Computational Physics Communications* 91(1-3):43-56. http://folding.bmc.uu.se/ http://www.gromacs.org.

Berg, O G; Kurland, C G. 2000 Jun. Why mitochondrial genes are most often found in nuclei. *Molecular Biology and Evolution* 17(6):951-961. http://mbe.oupjournals.org/cgi/content/full/17/6/951.

Berger, B; Leighton, T. 1998 Spring. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *Journal of Computational Biology* 5(1):27-40.

Berman, H M; Westbrook, J; Feng, Z; Gilliland, G; Bhat, T N; Weissig, H; Shindyalov, I N; Bourne, P E. 2000 Jan 1. The Protein Data Bank. *Nucleic Acids Research* 28(1):235-242. http://nar.oxfordjournals.org/cgi/content/full/28/1/235.

Bestor, T H. 2000 Oct 1. The DNA methyltransferases of mammals. *Human Molecular Genetics* 9(16):2395-2402.

Betancourt, M R; Thirumalai, D. 2002 Jan 24. Protein sequence design by energy landscaping. *Journal of Physical Chemistry B* 106(3):599-609.

Bienkowska, J R; Yu, L H; Zarakhovich, S; Rogers, R G, Jr; Smith, T F. 2000 Aug 15. Protein fold recognition by total alignment probability. *PROTEINS: Structure, Function, and Genetics* 40(3):451-462.

Bininda-Emonds, O R P; Gittleman, J L; Purvis, A. 1999 May. Building large trees by combining phylogenetic information: A complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews of the Cambridge Philosophical Society* 74(2):143-175.

Bininda-Emonds, O R P; Gittleman, J L; Steel, M A. 2002. The (Super)tree of life: Procedures, problems, and prospects. pp 265-289 ed: Futuyma, D J; Shaffer, H B; Simberloff, D. Annual Review of Ecology and Systematics. Vol: 33. Annual Reviews (Palo Alto, CA).

Biology, C f S. 2006 Dec 12. VOLUME. Yale University. http://www.csb.yale.edu/userguides/datamanip/volume/volume_descrip.html.

Bischoff, J; Domrachev, M; Federhen, S; Hotton, C; Leipe, D D; Soussov, V; Sternberg, R; Turner, S. 2004. NCBI Taxonomy. National Library of Medicine. ftp://ftp.ncbi.nih.gov/pub/taxonomy/.

Blakley, R L; Sorrentino, B P. 1998. *In vitro* mutations in dihydrofolate reductase that confer resistance to methotrexate: Potential for clinical application. *Human Mutation* 11(4):259-263.

Bleasby, A. 2000. needle, EMBOSS 3.0.0. European Bioinformatics Institute (Cambridge). http://emboss.sourceforge.net/.

Blouin, C; Butt, D; Roger, A J. 2005 Mar. Impact of taxon sampling on the estimation of rates of evolution at sites. *Molecular Biology and Evolution* 22(3):784-791. http://mbe.oxfordjournals.org/cgi/content/full/22/3/784.

Blundell, T L. 1991. Comparative analysis of protein three-dimensional structures and an approach to the inverse folding problem. *Symposium on Protein Conformation* 161:28-51. Chichester. Ciba Foundation Symposium. John Wiley and Sons.

Boden, M; Yuan, Z; Bailey, T L. 2006 Feb 14. Prediction of protein continuum secondary structure with probabilistic models based on NMR solved structures. *BioMedCentral Bioinformatics* 7:68. http://www.biomedcentral.com/1471-2105/7/68.

Boeckmann, B; Bairoch, A; Apweiler, R; Blatter, M-C; Estreicher, A; Gasteiger, E; Martin, M J; Michoud, K; O'Donovan, C; Phan, I; Pilbout, S; Schneider, M. 2003 Jan. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* 31(1):365-370. http://nar.oxfordjournals.org/cgi/content/full/31/1/365.

Bonneau, R; Baker, D. 2001. *Ab initio* protein structure prediction: Progress and prospects. pp 173-189 ed: Stroud, R M; Olson, W K; Sheetz, M P. Annual Review of Biophysics and Biomolecular Structure. Vol: 30. Annual Reviews (Palo Alto, CA).

Bowie, J U; Luthy, R; Eisenberg, D S. 1991 Jul 12. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253(5016):164-170.

Brenner, S E; Koehl, P; Levitt, M. 2000 Jan 1. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Research* 28(1):254-256. http://astral.berkeley.edu/ http://nar.oxfordjournals.org/cgi/content/full/28/1/254.

Brenner, S E; Walker, N; Koehl, P; Levitt, M. 2000 Sep 25. ASTRAL home page: The ASTRAL compendium for sequence and structure analysis. University of California. http://astral.berkeley.edu.

Brophy, V H; Vasquez, J R; Nelson, R G; Forney, J R; Rosowsky, A; Sibley, C H. 2000 Apr. Identification of *Cryptosporidium parvum* dihydrofolate reductase inhibitors by complementation in *Saccharomyces cerevisiae*. *Antimicrobial Agents and Chemotherapy* 44(4):1019-1028.

Brower, A V Z; DeSalle, R; Vogler, A. 1996. Gene trees, species trees, and systematics: A cladistic perspective. pp 423-450 ed: Fautin, D G; Futuyma, D J; James, F C. Annual Review of Ecology and Systematics. Vol: 27. Annual Reviews (Palo Alto, CA).

Brown, C J; Takayama, S; Campen, A M; Vise, P; Marshall, T W; Oldfield, C J; Williams, C J; Dunker, A K. 2002 Jul. Evolutionary rate heterogeneity in proteins with long disordered regions. *Journal of Molecular Evolution* 55(1):104-110.

Brown, J R; Doolittle, W F. 1997 Dec. Archaea and the procaryote-to-eucaryote transition. *Microbiology and Molecular Biology Reviews* 61(4):456-502. http://mmbr.asm.org/cgi/content/abstract/61/4/456.

Brown, W M; Prager, E M; Wang, A; Wilson, A C. 1982 Jul. Mitochondrial DNA sequences of primates: Tempo and mode of evolution. *Journal of Molecular Evolution* 18(4):225-239. http://hdl.handle.net/2027.42/48036.

Bryant, S H. 2004. PDBeast. National Library of Medicine. http://www.ncbi.nlm.nih.gov/Structure/PDBEAST/pdbeast.shtml.

Buhler, R; Hempel, J; Kaiser, R; Von Wartburg, J-P; Vallee, B L; Jornvall, H. 1984 Oct 15. Human alcohol dehydrogenase: Structural differences between the beta and gamma subunits suggest parallel duplications in isoenzyme evolution and predominant expression of separate gene descendants in livers of different mammals. *Proceedings of the National Academy of Sciences USA* 81(20):6320-6324. http://www.pnas.org/cgi/content/abstract/81/20/6320.

Bull, J J; Huelsenbeck, J P; Cunningham, C W; Swofford, D L; Waddell, P J. 1993 Sep. Partitioning and combining data in phylogenetic analysis. *Systematic Biology* 42(3):384-397.

Burke, D F; Deane, C M; Nagarajaram, H A; Campillo, N; Martin-Martinez, M; Mendes, J; Molina, F; Perry, J; Reddy, B V B; Soares, C M; Steward, R E; Williams, M; Carrondo, M A; Blundell, T L; Mizuguchi, K. 1999. An iterative structure-assisted approach to sequence alignment and comparative modeling. *PROTEINS: Structure, Function, and Genetics* Suppl 3:55-60.

de Castro, E; Sigrist, C J A; Gattiker, A; Bulliard, V; Langendijk-Genevaux, P S; Gasteiger, E; Bairoch, A; Hulo, N. 2006. ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research* 34(Suppl 2):W362-W365. http://nar.oxfordjournals.org/cgi/content/full/34/suppl_2/W362.

Cavalli-Sforza, L L; Edwards, A W F. 1967 May. Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics* 19(3 pt 1):233-257. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=6026583.

Chakrabarti, P; Pal, D. 2001. The interrelationships of side-chain and main-chain conformations in proteins. *Progress in Biophysics & Molecular Biology* 76(1-2):1-102. http://www.serc.iisc.ernet.in/~dpal/publications.html.

Chandrasekharan, U M; Sanker, S; Glynias, M J; Karnik, S S; Husain, A. 1996 Jan 26. Angiotensin II-forming activity in a reconstructed ancestral chymase. *Science* 271(5248):502-505.

Chang, B S W; Campbell, D L. 2000 Aug. Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences. *Molecular Biology and Evolution* 17(8):1220-1231. http://mbe.oupjournals.org/cgi/content/full/17/8/1220.

Chang, B S W; Donoghue, M J. 2000 Mar 1. Recreating ancestral proteins. *Trends in Ecology and Evolution* 15(3):109-114.

Chang, M S S; Benner, S A. 2004 Aug 6. Empirical analysis of protein insertions and deletions: Determining parameters for the correct placement of gaps in protein sequence alignments. *Journal of Molecular Biology* 341(2):617-631.

Chase, T, Jr. 2005. YAHK is a cinnamyl alcohol dehydrogenase. Oral Communication. To: Smith, A W (New Brunswick, NJ).

Chater, K F; Horinouchi, S. 2003 Apr. Signalling early developmental events in two highly diverged *Streptomyces* species. *Molecular Microbiology* 48(1):9-15.

Chen, J W; Romero, P R; Uversky, V N; Dunker, A K. 2006 Apr. Conservation of intrinsic disorder in protein domains and families II: Functions of conserved disorder. *Journal of Proteome Research* 5(4):888-898.

Cheunq, B; Holmes, R S; Easteal, S; Beacham, I R. 1999 Jan. Evolution of class I alcohol dehydrogenase genes in catarrhine primates: Gene conversion, substitution rates, and gene regulation. *Molecular Biology and Evolution* 16(1):23-36. http://mbe.oxfordjournals.org/cgi/content/abstract/16/1/23.

Chung, S Y; Subbiah, S. 1996 Oct 15. A structural explanation for the twilight zone of protein sequence homology. *Structure with Folding & Design* 4(10):1123-1127.

Ciccarelli, F D; Doerks, T; von Mering, C; Creevey, C J; Snel, B; Bork, P. 2006 Mar 3. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311(5765):1283-1287.

Clark, D E; Westhead, D R. 1996 Aug. Evolutionary algorithms in computer-aided molecular design. *Journal of Computer-Aided Molecular Design* 10(4):337-358.

Coar, K. 2006 Jul 7. The open source definition. Open Source Initiative. http://www.opensource.org/docs/osd.

Cochran, D A E; Penel, S; Doig, A J. 2001 Mar. Effect of the N1 residue on the stability of the alpha-helix for all 20 amino acids. *Protein Science* 10(3):463-470.

Cody, V; Chan, D; Galitsky, N; Rak, D; Luft, J R; Pangborn, W; Queener, S F; Laughton, C A; Stevens, M F. 2000 Apr 4. Structural studies on bioactive compounds 30: Crystal structure and molecular modeling studies on the *Pneumocystis carinii* dihydrofolate reductase cofactor complex with TAB, a highly selective antifolate. *Biochemistry* 39(13):3556-3564.

Coeytaux, K; Poupon, A. 2005 May 1. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics* 21(9):1891-1900. http://bioinformatics.oxfordjournals.org/cgi/content/full/21/9/1891.

Collins, T M; Wimberger, P H; Naylor, G J P. 1994. Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Systematic Biology* 43(4):482-496.

Colloc'h, N; Etchebest, C; Thoreau, E; Henrissat, B; Mornon, J P. 1993 Jun. Comparison of three algorithms for the assignment of secondary structure in proteins: The advantage of a consensus assignment. *Protein Engineering* 6(4):377-382.

Commons, C. 2006. Creative Commons Attribution-Share Alike 2.5 Generic License. Creative Commons. http://creativecommons.org/licenses/by-sa/2.5/.

Commons, S. 2007. Databases and Creative Commons. MIT CSAIL. http://sciencecommons.org/resources/faq/databases/.

Cootes, A P; Curmi, P M G; Cunningham, R; Donnelly, C A; Torda, A E. 1998 Aug 1. The dependence of amino acid pair correlations on structural environment. *PROTEINS: Structure, Function, and Genetics* 32(2):175-189.

Corana, A; Marchesi, M; Martini, C; Ridella, S. 1987 Sep. Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. *ACM Transactions on Mathematical Software* 13(3):262-280.

Cotter, P J; Caffrey, D R; Shields, D C. 2002 Jan 1. Improved database searches for orthologous sequences by conditioning on outgroup sequences. *Bioinformatics* 18(1):83-91. http://bioinformatics.oupjournals.org/cgi/content/abstract/18/1/83.

Coutinho, P M; Henrissat, B. 1999. Carbohydrate-active enzymes: An integrated database approach. pp 3-12 in *Recent Advances in Carbohydrate Bioengineering* ed: Gilbert, H J; Davies, G J; Henrissat, B; Svensson, B. The Royal Society of Chemistry (Cambridge).

Coutinho, P M; Henrissat, B. 2007. Carbohydrate Active Enzymes database. UMR6098, CNRS/Université de Provence/Université de la Méditerranée. http://www.cazy.org.

Creamer, T P; Srinivasan, R; Rose, G D. 1995 Dec 19. Modeling unfolded states of peptides and proteins. *Biochemistry* 34(50):16245-16250.

Crescenzi, P; Goldman, D; Papadimitriou, C H; Piccolboni, A; Yannakakis, M. 1998 Fall. On the complexity of protein folding. *Journal of Computational Biology* 5(3):423-465.

Cuff, J A; Barton, G J. 1999 Mar 1. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *PROTEINS: Structure, Function, and Genetics* 34(4):508-519.

Cuff, J A; Barton, G J. 2000 Aug 15. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *PROTEINS: Structure, Function, and Genetics* 40(3):502-511.

Cui, W; DeWitt, J G; Miller, S M; Wu, W. 1999 May 27. No metal cofactor in orotidine 5'-monophosphate decarboxylase. *Biochemical & Biophysical Research Communications* 259(1):133-135.

Cummings, M P; Otto, S P; Wakeley, J. 1995 Sep. Sampling properties of DNA sequence data in phylogenetic analysis. *Molecular Biology and Evolution* 12(5):814-822. http://mbe.oupjournals.org/cgi/content/abstract/12/5/814.

Cummings, M P; Otto, S P; Wakeley, J. 1999 Jun. Genes and other samples of DNA sequence data for phylogenetic inference. *Biological Bulletin* 196(3):345-350. http://www.biolbull.org/cgi/reprint/196/3/345.

Cunningham, C W; Omland, K E; Oakley, T H. 1998 Sep. Reconstructing ancestral character states: A critical reappraisal. *Trends in Ecology and Evolution* 13(9):361-366.

Cunningham, C W; Zhu, H; Hillis, D M. 1998 Aug. Best-fit maximum-likelihood models for phylogenetic inference: Empirical tests with known phylogenies. *Evolution* 52(4):978-987.

D'Alfonso, G; Tramontano, A; Lahm, A. 2001 May. Structural conservation in single-domain proteins: Implications for homology modeling. *Journal of Structural Biology* 134(2-3):246-256.

Dacks, J B; Marinets, A; Doolittle, W F; Cavalier-Smith, T; Logsdon, J M, Jr. 2002 Jun 1. Analyses of RNA Polymerase II genes from free-living protists: Phylogeny, long branch attraction, and the eukaryotic big bang. *Molecular Biology and Evolution* 19(6):830-840. http://mbe.oupjournals.org/cgi/content/full/19/6/830.

Dalal, S; Balasubramanian, S; Regan, L. 1997 Jul. Protein alchemy: Changing ß-sheet into alpha-helix. *Nature Structural Biology* 4(7):548-552.

Dalton, J A R; Jackson, R M. 2007 Aug. An evaluation of automated homology modelling methods at low target–template sequence similarity. *Bioinformatics* 23(15):1901-1908. http://bioinformatics.oxfordjournals.org/cgi/content/full/23/15/1901.

Das, B; Meirovich, H. 2001 May 15. Optimization of solvation models for predicting the structure of surface loops in proteins. *PROTEINS: Structure, Function, and Genetics* 43(3):303-314.

Davis, I W; Chen, V B; Immormino, R M; Arendall, W B, III. 2007. MolProbity, 3.13. Duke University. http://molprobity.biochem.duke.edu.

Dayhoff, M O; Schwartz, R M; Orcutt, B C. 1978. A model of evolutionary change in proteins. pp 345-352 in *Atlas of Protein Sequence and Structure* ed: Dayhoff, M O. Vol: 5 (Suppl 3). National Biomedical Research Foundation (Washington, DC). http://www.dayhoff.cc/.

Dean, A M; Golding, G B. 1997 Apr 1. Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase. *Proceedings of the National Academy of Sciences USA* 94(7):3104-3109. http://www.pnas.org/cgi/content/abstract/94/7/3104.

Dean, A M. 1998 Jan/Feb. The molecular anatomy of an ancient adaptive event. *American Scientist* 86(1):26-37.

Deane, C M; Blundell, T L. 2001 Mar 1. CODA: A combined algorithm for predicting the structurally variable regions of protein models. *Protein Science* 10(3):599-612. http://www.proteinscience.org/cgi/content/full/10/3/599.

Defay, T; Cohen, F E. 1995 Nov. Evaluation of current techniques for *ab-initio* protein structure prediction. *PROTEINS: Structure, Function, and Genetics* 23(3):431-445.

Degan, P; Carpano, P; Cercignani, G; Montagnoli, G. 1989 Mar. A fluorescence study of substrate and inhibitor binding to bovine liver dihydrofolate reductase. *International Journal of Biochemistry* 21(3):291-295.

Desjarlais, J R; Handel, T M. 1995 Aug. New strategies in protein design. *Current Opinion in Biotechnology* 6(4):460-466.

Desjarlais, J R; Clarke, N D. 1998 Aug. Computer search algorithms in protein modification and design. *Current Opinion in Structural Biology* 8(4):471-475.

Desmet, J; Spriet, J; Lasters, I. 2002 Jul 1. Fast and Accurate Side-chain Topology and Energy Refinement (FASTER) as a new method for protein structure optimization. *PROTEINS: Structure, Function, and Genetics* 48(1):31-43.

Dobzhansky, T. 1973 Mar. Nothing in biology makes sense except in the light of evolution. *American Biology Teacher* 35:125-129.

Dollo, L. 1893. Les lois de l'evolution. *Bulletin de la Societe Belge de Geologie, de Paleontologie, et d'Hydrologie* 7:164-166.

Domingues, F S; Lackner, P; Andreeva, A; Sippl, M J. 2000 Apr 7. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *Journal of Molecular Biology* 297(4):1003-1013. http://lore.came.sbg.ac.at/Publications/publications.html.

Drennan, D; Richards, F M; Kahn, P C. 1993. DSSP modification. Rutgers University. http://cesario.rutgers.edu/~mallows/mDSSP.htm.

Drennan, D. 2001. *The Geometrical Analysis of the Structure of Proteins (GASP) with implications for protein folding.* Dissertation. Rutgers University, Microbiology and Molecular Genetics Graduate Program (New Brunswick, NJ): Department of Biochemistry and Microbiology. http://cesario.rutgers.edu/~mallows/abstract1.html http://cesario.rutgers.edu/~mallows/resume.html.

Duffy, T H; Beckman, S B; Peterson, S M; Vitols, K S; Huennekens, F M. 1987 May 25. L1210 dihydrofolate reductase: Kinetics and mechanism of activation by various agents. *Journal of Biological Chemistry* 262(15):7028-7033. http://www.jbc.org/cgi/content/abstract/262/15/7028.

Durbin, R M; Eddy, S R; Krogh, A; Mitchison, G J. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press (Cambridge).

Dutheil, J; Galtier, N. 2007 Nov 30. Detecting groups of co-evolving positions in a molecule: A clustering approach. *BiomedCentral Evolutionary Biology* 7(1):242. http://www.biomedcentral.com/1471-2148/7/242.

Easteal, S. 1990 Jan. The pattern of mammalian evolution and the relative rate of molecular evolution. *Genetics* 124(1):165-173. http://www.genetics.org/cgi/content/abstract/124/1/165.

Eck, R V; Dayhoff, M O: ed. 1966. *Atlas of Protein Sequence and Structure*. Volume 2. National Biomedical Research Foundation (Silver Spring, Maryland).

Eddy, S R. 1995. Multiple alignment using hidden Markov models. *International Conference on Intelligent Systems for Molecular Biology* 3:114-120. Menlo Park, CA. AAAI/MIT Press.

Eddy, S R. 1998 Oct 1. Profile hidden Markov models. *Bioinformatics* 14(9):755-763. ftp://ftp.genetics.wustl.edu/pub/eddy/papers/ http://bioinformatics.oupjournals.org/cgi/content/abstract/14/9/755.

Eddy, S R; Birney, E. 2003. HMMer, 2.3.2. HHMI/Washington University School of Medicine (Saint Louis, Missouri). http://hmmer.janelia.org/.

Edgar, R C; Sjolander, K. 2003. Simultaneous sequence alignment and tree construction using hidden Markov models. *Pacific Symposium on Biocomputing* 8:180-191.

Edwards, A W F; Cavalli-Sforza, L L. 1964. Reconstruction of evolutionary trees. pp 67-76 in *Phenetic and Phylogenetic Classification* ed: Heywood, V H; McNeill, J. Vol: 6. Systematics Association (London).

Edwards, R J; Shields, D C. 2004 Sep 6. GASP: Gapped Ancestral Sequence Prediction for proteins. *BioMedCentral Bioinformatics* 5:123. http://www.biomedcentral.com/1471-2105/5/123.

Efron, B. 1979 Jan. Bootstrap methods: Another look at the jacknife. *Annals of Statistics* 7(1):1-26.

Eisenhaber, F; Persson, B; Argos, P. 1995. Protein structure prediction: Recognition of primary, secondary, and tertiary structural features from amino-acid sequence. *Critical Reviews in Biochemistry and Molecular Biology* 30(1):1-94.

Elofsson, A. 2002 Feb 15. A study on protein sequence alignment quality. *PROTEINS: Structure, Function, and Genetics* 46(3):330-339.

Embley, T M; Stackebrandt, E. 1994. The molecular phylogeny and systematics of the actinomycetes. pp 257-289 ed: Ornson, L N; Balows, A; Greenberg, E P. Annual Review of Microbiology. Vol: 48.

Engel, L. 2007 Dec. Loop search results. Oral and electronic communication. To: Smith, A W (New Brunswick, NJ).

Eswar, N; Ramakrishnan, C. 2000 Apr. Deterministic features of side-chain—main-chain hydrogen bonds in globular protein structures. *Protein Engineering* 13(4):227-238. http://peds.oxfordjournals.org/cgi/content/full/13/4/227.

Eswar, N; Ramakrishnan, C; Srinivasan, N. 2003 May 1. Stranded in isolation: Structural role of isolated extended strands in proteins. *Protein Engineering* 16(5):331-339. http://peds.oupjournals.org/cgi/content/full/16/5/331.

Falicov, A; Cohen, F E. 1996 May 24. A surface of minimum area metric for the structural comparison of proteins. *Journal of Molecular Biology* 258(5):871-892.

Fariselli, P; Olmea, O; Valencia, A; Casadio, R. 2001 Nov. Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering* 14(11):835-843. http://peds.oxfordjournals.org/cgi/content/full/14/11/835.

Farnum, M F; Magde, D; Howell, E E; Hirai, J T; Warren, M S; Grimsley, J K; Kraut, J. 1991 Dec 10. Analysis of hydride transfer and cofactor fluorescence decay in mutants of dihydrofolate

reductase: Possible evidence for participation of enzyme molecular motions in catalysis. *Biochemistry* 30(49):11567-11579.

Farris, J S. 1977 Mar. Phylogenetic analysis under Dollo's Law. *Systematic Zoology* 26(1):77-88.

Farris, J S. 1983. The logical basis of phylogenetic analysis. pp 7-36 in *Proceedings of the Second Meeting of the Willi Hennig Society* ed: Platnick, N I; Funk, V A. Advances in Cladistics. Vol: 2. Columbia University Press (New York).

Felsenstein, J. 1973 Sep. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics* 25(5):471-492. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=4741844.

Felsenstein, J. 1978 Dec. Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology* 27(4):401-410.

Felsenstein, J. 1981 Nov. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological Journal of the Linnean Society* 16(3):183-196.

Felsenstein, J. 1984a Jan. Distance methods for inferring phylogenies: A justification. *Evolution* 38(1):16-24.

Felsenstein, J. 1984b. The statistical approach to inferring evolutionary trees and what it tells us about parsimony and compatibility. pp 169-191 in *Cladistics: Perspectives in the Reconstruction of Evolutionary History* ed: Duncan, T; Stuessy, T F. Columbia University Press (New York).

Felsenstein, J. 1985a Jul. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39(4):783-791.

Felsenstein, J. 1985b Jan. Phylogenies and the comparative method. *American Naturalist* 125(1):1-25.

Felsenstein, J. 1988. Phylogenies from molecular sequences: Inference and reliability. pp 521-565 ed: Campbell, A; Baker, B S; Herskowitz, I. Annual Review of Genetics. Vol: 22. Annual Reviews (Palo Alto, CA).

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package), 3.5c. Department of Genetics, University of Washington (Seattle). http://evolution.genetics.washington.edu/phylip.html.

Fiser, A; Do, R K G; Sali, A. 2000 Sep. Modeling of loops in protein structures. *Protein Science* 9(9):1753-1773. http://www.proteinscience.org/cgi/content/full/9/9/1753.

Fisher, H F; Conn, E E; Vennesland, B; Westheimer, F H. 1953 Jun 1. The enzymatic transfer of hydrogen I: The reaction catalyzed by alcohol dehydrogenase. *Journal of Biological Chemistry* 202(2):687-697. http://www.jbc.org/cgi/reprint/202/2/687.

Fitch, W M; Margoliash, E. 1967 Jan 20. Construction of phylogenetic trees. *Science* 155:279-284.

Fitzpatrick, D A; Logue, M E; Stajich, J E; Butler, G. 2006 Nov 22. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BioMedCentral Bioinformatics* 6:99. http://www.biomedcentral.com/1471-2148/6/99.

Flohil, J A; Vriend, G; Berendsen, H J C. 2002 Sep 1. Completion and refinement of 3-D homology models with restricted molecular dynamics: Application to targets 47, 58, and 111 in

the CASP modeling competition and posterior analysis. *PROTEINS: Structure, Function, and Genetics* 48(4):593-604.

Fornasari, M S; Parisi, G; Echave, J. 2002 Mar 1. Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. *Molecular Biology and Evolution* 19(3):352-356. http://mbe.oupjournals.org/cgi/content/full/19/3/352.

Foundation, F S. 2002. Gnu Compiler Collection, 3.2.3. Free Software Foundation (Boston, MA). http://gcc.gnu.org/onlinedocs/gcc-3.2.3/gcc/.

Foundation, F S. 2007 Nov 19. Licenses. Free Software Foundation (Boston, MA). http://www.gnu.org/licenses/.

Frigo, M; Johnson, S G. 2005 Feb. The design and implementation of FFTW3. *Proceedings of the IEEE* 93(2):216-231. http://www.fftw.org.

Fuellen, G; Wagele, J-W; Giegerich, R. 2001 Dec 1. Minimum conflict: A divide-and-conquer approach to phylogeny estimation. *Bioinformatics* 17(12):1168-1191. http://bioinformatics.oupjournals.org/cgi/content/abstract/17/12/1168.

Fukami-Kobayashi, K; Schreiber, D R; Benner, S A. 2002 Jun 7. Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. *Journal of Molecular Biology* 319(3):729-743.

Futuyma, D J. 1986. *Evolutionary Biology*. Edition: 2nd. Sinauer Associates (Sunderland, MA). ISBN 0-87893-188-0.

Galtier, N. 2001 May. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution* 18(5):866-873. http://mbe.oupjournals.org/cgi/content/full/18/5/866.

Galtier, N. 2004 Feb. Sampling properties of the bootstrap support in molecular phylogeny: Influence of nonindependence among sites. *Systematic Biology* 53(1):38-46.

Garnier, J; Osguthorpe, D J; Robson, B. 1978 Mar 25. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology* 120(1):97-120.

Gatz, D F; Smith, L. 1995 Jun. The standard error of a weighted mean concentration: I. Bootstrapping vs other methods. *Atmospheric Environment* 29(11):1185-1193.

Gaucher, E A; Miyamoto, M M; Benner, S A. 2001 Jan 16. Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. *Proceedings of the National Academy of Sciences USA* 98(2):548-552. http://www.pnas.org/cgi/content/full/98/2/548.

Gaucher, E A; Das, U K; Miyamoto, M M; Benner, S A. 2002 Apr 1. The crystal structure of eEF1A refines the functional predictions of an evolutionary analysis of rate changes among elongation factors. *Molecular Biology and Evolution* 19(4):569-573. http://mbe.oupjournals.org/cgi/content/full/19/4/569.

Gaur, L K; Hughes, A L; Heise, E R; Gutknecht, J. 1992 Jul. Maintenance of DQβ1 polymorphisms in primates. *Molecular Biology and Evolution* 9(4):599-609. http://mbe.oupjournals.org/cgi/content/abstract/9/4/599.

Georis, J; Giannotta, F; Lamotte-Brasseur, J; Devreese, B; Van Beeumen, J; Granier, B; Frere, J-M. 1999 Sep 3. Sequence, overproduction and purification of the family 11 endo-beta-1,4-xylanase encoded by the xyl1 gene of *Streptomyces sp. S38*. *Gene* 237(1):123-133.

Gerrits, G P; Klaassen, C; Coenye, T; Vandamme, P; Meis, J F. 2005 Jul. *Burkholderia fungorum* septicemia. *Emerging Infectious Diseases* 11(7). http://www.cdc.gov/ncidod/EID/vol11no07/04-1290.htm.

Gerstein, M; Lynden-Bell, R M. 1993 Mar 20. What is the natural boundary of a protein in solution? *Journal of Molecular Biology* 230(2):641-650. http://papers.gersteinlab.org/e-print/prot-bound-jmb/preprint.pdf.

Gerstein, M; Tsai, J; Levitt, M. 1995 Jun 23. The volume of atoms on the protein surface: Calculated from simulation, using Voroni polyhedra. *Journal of Molecular Biology* 249(5):955-966. http://papers.gersteinlab.org/e-print/surfprotvol/preprint.pdf.

Gerstein, M; Levitt, M. 1996 Jun 12-15. Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *International Conference on Intelligent Systems for Molecular Biology* 4:59-67. St. Louis. AAAI/MIT Press. http://papers.gersteinlab.org/e-print/programming-alignment/preprint.pdf.

Gerstein, M; Levitt, M. 1998 Feb. Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. *Protein Science* 7(2):445-456. http://www.proteinscience.org/cgi/content/abstract/7/2/445 http://papers.gersteinlab.org/e-print/scop-str-prosci-reprint.pdf.

Gerstein, M; Richards, F M. 2001. Protein geometry: Distances, areas, and volumes. pp 531-539 in *Crystallography of biological macromolecules* ed: Arnold, E; Rossmann, M G. International Tables for Crystallography. Vol: F. Kluwer (Dordrecht, Netherlands). http://bioinfo.mbb.yale.edu/papers/.

Gertz, C M. 2006. BLAST. National Library of Medicine. http://www.ncbi.nlm.nih.gov/blast/Blast.cgi.

Ghim, S Y; Nielsen, P; Neuhard, J. 1994 Mar. Molecular characterization of pyrimidine biosynthesis genes from the thermophile *Bacillus caldolyticus*. *Microbiology* 140(3):479-491.

Gibb, G C; Kardailsky, O; Kimball, R T; Braun, E L; Penny, D. 2007 Jan. Mitochondrial genomes and avian phylogeny: Complex characters and resolvability without explosive radiations. *Molecular Biology and Evolution* 24(1):269-280. http://mbe.oxfordjournals.org/cgi/content/full/24/1/269.

Gibbs, S; Collard, M; Wood, B. 2000 Sep 26. Soft-tissue characters in higher primate phylogenetics. *Proceedings of the National Academy of Sciences USA* 97(20):11130-11132. http://www.pnas.org/cgi/content/full/97/20/11130.

Gilquin, B; Racape, J; Wrisch, A; Visan, V; Lecoq, A; Grissmer, S; Menez, A; Gasparini, S. 2002 Oct 4. Structure of the BgK-Kv1.1 complex based on distance restraints identified by double mutant cycles: Molecular basis for convergent evolution of Kv1 channel blockers. *Journal of Biological Chemistry* 277(40):37406-37413. http://www.jbc.org/cgi/content/full/277/40/37406.

Glazko, G V; Nei, M. 2003 Mar 1. Estimation of divergence times for major lineages of primate species. *Molecular Biology and Evolution* 20(3):424-434. http://mbe.oupjournals.org/cgi/content/full/20/3/424.

Gobel, U; Sander, C; Schneider, R; Valencia, A. 1994 Apr. Correlated mutations and residue contact prediction. *PROTEINS: Structure, Function, and Genetics* 18(4):309-317.

Godzik, A. 1996 Jul. The structural alignment between two proteins: Is there a unique answer? *Protein Science* 5(7):1325-1338.

Gogarten, J P; Doolittle, W F; Lawrence, J G. 2002 Dec 1. Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution* 19(12):2226-2238. http://mbe.oxfordjournals.org/cgi/content/full/19/12/2226.

Gogarten, J P; Townsend, J P. 2005 Sep. Horizontal gene transfer, genome innovation, and evolution. *Nature Reviews Microbiology* 3(9):679-687. http://dx.doi.org/10.1038/nrmicro1204.

Golding, G B; Dean, A M. 1998 Apr. The structural basis of molecular adaptation. *Molecular Biology and Evolution* 15(4):355-369. http://mbe.oupjournals.org/cgi/content/abstract/15/4/355.

Goldman, N; Yang, Z. 1994 Sep. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* 11(5):725-736. http://abacus.gene.ucl.ac.uk/ziheng/cv.html http://mbe.oupjournals.org/cgi/content/abstract/11/5/725.

Goldman, N; Thorne, J L; Jones, D T. 1996 Oct 25. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analysis. *Journal of Molecular Biology* 263(2):196-208.

Goldman, N; Thorne, J L; Jones, D T. 1998 May 1. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149(1):445-458. http://www.genetics.org/cgi/content/full/149/1/445.

Goldsmith-Fischman, S; Honig, B. 2003 Sep. Structural genomics: Computational methods for structure analysis. *Protein Science* 12(9):1813-1821. http://www.proteinscience.org/cgi/content/full/12/9/1813.

Golubchik, T; Wise, M J; Easteal, S; Jermiin, L S. 2007 Nov 1. Mind the gaps: Evidence of bias in estimates of multiple sequence alignments. *Molecular Biology and Evolution* 24(11):2433-2442.

Gomes, A C; Miranda, I; Silva, R M; Moura, G R; Thomas, B; Akoulitchev, A; Santos, M A S. 2007. A genetic code alteration generates a proteome of high diversity in the human pathogen *Candida albicans*. *Genome Biology* 8(10):R206. http://genomebiology.com/2007/8/10/R206.

Gonnet, G H; Cohen, M A; Benner, S A. 1992 Sep 18. Exhaustive matching of the entire protein sequence database. *Science* 256(5077):1433-1445.

Goodsell, D S. 1999 Sep. The molecular perspective: Methotrexate. *Stem Cells* 17(5):314-315. http://stemcells.alphamedpress.org/cgi/content/full/17/5/314.

Goonesekere, N C W; Lee, B-K. 2004 May. Frequency of gaps observed in a structurally aligned protein pair database suggests a simple gap penalty function. *Nucleic Acids Research* 32(9):2838-2843. http://nar.oxfordjournals.org/cgi/content/full/32/9/2838.

Gophna, U; Doolittle, W F; Charlebois, R L. 2005 Feb 15. Weighted genome trees: Refinements and applications. *Journal of Bacteriology* 187(4):1305-1316. http://jb.asm.org/cgi/content/full/187/4/1305 http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=15687194.

Gordon, J; Sibley, L D. 2005. Comparative genome analysis reveals a conserved family of actin-like proteins in apicomplexan parasites. *BioMedCentral Genomics* 6(1):179. http://www.biomedcentral.com/1471-2164/6/179.

Graham, S W; Olmstead, R G; Barrett, S C H. 2002 Oct 1. Rooting phylogenetic trees with distant outgroups: A case study from the commelinoid monocots. *Molecular Biology and Evolution* 19(10):1769-1791. http://mbe.oupjournals.org/cgi/content/abstract/19/10/1769.

Gregson, A; Plowe, C V. 2005 Mar. Mechanisms of resistance of malaria parasites to antifolates. *Pharmacological Reviews* 57(1):117-145. http://pharmrev.aspetjournals.org/cgi/content/full/57/1/117.

de Groot, B. 2004 Oct 15. [gmx-users] pdb2gmx and hydrogen nomenclature. http://www.gromacs.org/pipermail/gmx-users/2004-October/012679.html.

Gunasekaran, K; Nagarajaram, H A; Ramakrishnan, C; Balaram, P. 1998 Feb 6. Stereochemical punctuation marks in protein structures: Glycine and proline containing helix stop signals. *Journal of Molecular Biology* 275(5):917-932.

van Gunsteren, W F; Billeter, S R; Eising, A A; Hunenberger, P H; Kruger, P; Mark, A E; Scott, W R P; Tironi, I G. 1996. *Biomolecular simulation: The GROMOS96 manual and user guide*. Vdf Hochschulverlag ETHZ (Zurich, Switzerland).

Gupta, R S. 1998 Dec. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eucaryotes. *Microbiology and Molecular Biology Reviews* 62(4):1435-1491. http://mmbr.asm.org/cgi/content/abstract/62/4/1435.

Gupta, R S. 2000 Oct. The phylogeny of proteobacteria: Relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiology Reviews* 24(4):367-402.

Gupta, R S. 2001 Dec. The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. *International Microbiology* 4(4):187-202.

Gupta, R S. 2005. Molecular sequences and the early history of life. in *Microbial Phylogeny and Evolution: Concepts and Controversies* ed: Sapp, J. Oxford University Press (New York). http://www.bacterialphylogeny.info/bacteria.html.

Gupta, R S. 2007 May 28. Bacterial (Prokaryotic) Phylogeny. http://www.bacterialphylogeny.info/bacteria.html http://www.bacterialphylogeny.com/index.html.

Gutell, R R; Larsen, N; Woese, C R. 1994 Mar. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiological Reviews* 58(1):10-26.

von Haeseler, A; Churchill, G A. 1993 Jul. Network models for sequence evolution. *Journal of Molecular Evolution* 37(1):77-85.

von Haeseler, A; Strimmer, K S. 2003. Phylogeny inference based on maximum-likelihood methods with TREE-PUZZLE. in *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny* ed: Salemi, M; Vandamme, A-M. Cambridge University Press (Cambridge). http://www.stat.uni-muenchen.de/~strimmer/cv.html.

Halpern, A L; Bruno, W J. 1998 Jul. Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. *Molecular Biology and Evolution* 15(7):910-917. http://mbe.oupjournals.org/cgi/content/abstract/15/7/910.

Hancock, J M; Dover, G A. 1990 Oct 25. 'Compensatory slippage' in the evolution of ribosomal RNA genes. *Nucleic Acids Research* 18(20):5949-5954. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=2235480.

Hankeln, T; Klawitter, S; Kramer, M; Burmester, T. 2006 Jul. Molecular characterization of hemoglobin from the honeybee *Apis mellifera*. *Journal of Insect Physiology* 52(7):701-710.

Harris, P; Poulsen, J-C N; Jensen, K F; Larsen, S. 2000 Apr 18. Structural basis for the catalytic mechanism of a proficient enzyme: Orotidine 5'-monophosphate decarboxylase. *Biochemistry* 39(15):4217-4224.

Hartigan, J A. 1973 Mar. Minimum mutation fits to a given tree. *Biometrics* 29(1):53-65.

Hasegawa, M; Fujiwara, M. 1993 Mar. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Molecular Phylogenetics and Evolution* 2(1):1-5.

Hastings, W K. 1970 Apr. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1):97-109. http://en.wikipedia.org/wiki/Metropolis-Hastings_algorithm.

Havel, T F. 1998. Distance geometry: Theory, algorithms, and chemical applications. in *Molecular Mechanics* ed: Kollman, P; Allinger, N. Encyclopedia of Computational Chemistry. Vol: 5. Wiley (Bern, Switzerland). http://web.mit.edu/tfhavel/www/Public/dg-review.pdf.

Havel, T F. 2007 Dec 12. Home Page for Timothy F. Havel. MIT. http://web.mit.edu/tfhavel/www/.

Hayes, R J; Bentzien, J; Ary, M L; Hwang, M Y; Jacinto, J M; Vielmetter, J; Kundu, A; Dahiyat, B I. 2002 Dec 10. Combining computational and experimental screening for rapid optimization of protein properties. *Proceedings of the National Academy of Sciences USA* 99(25):15926-15931. http://www.pnas.org/cgi/content/full/99/25/15926.

Heckert, N A; Filliben, J J. 2003. Weighvar. in *Dataplot Reference Manual*. NIST Handbook 148. Vol: 2. National Institute of Standards and Technology (Gaithersburg, Maryland). http://www.itl.nist.gov/div898/software/dataplot/refman2/ch2/weighvar.pdf.

Henikoff, S; Henikoff, J G. 1992 Nov 1. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences USA* 89(22):10915-10919. http://www.pnas.org/cgi/content/abstract/89/22/10915.

Henikoff, S; Henikoff, J G. 2000. Amino acid substitution matrices. pp 73-97 in *Analysis of Amino Acid Sequences* ed: Bork, P. Advances in Protein Chemistry. Vol: 54. Academic Press (San Diego).

Hennig, W. 1979. *Phylogenetic Systematics*. Edition: 2nd. University of Illinois Press (Urbana). ISBN 0-252-00745-X.

Henrissat, B. 1991 Dec 1. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochemical Journal* 280(2):309-316. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=1747104.

Henrissat, B; Bairoch, A. 1993 Aug 1. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochemical Journal* 293(3):781-788. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=8352747.

Henrissat, B; Callebaut, I; Fabrega, S; Lehn, P; Mornon, J P; Davies, G J. 1995 Jul 18. Conserved catalytic machinery and the prediction of a common fold for several families of

glycosyl hydrolases. *Proceedings of the National Academy of Sciences USA* 92(15):7090-7094. http://www.pnas.org/cgi/content/abstract/92/15/7090.

Henrissat, B; Davies, G J. 2000 Dec. Glycoside hydrolases and glycosyltransferases: Families, modules, and implications for genomics. *Plant Physiology* 124(4):1515-1519. http://www.plantphysiol.org/cgi/content/full/124/4/1515.

Hickson, R E; Simon, C; Cooper, A; Spicer, G S; Sullivan, J; Penny, D. 1996 Jan. Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. *Molecular Biology and Evolution* 13(1):150-169. http://mbe.oupjournals.org/cgi/content/abstract/13/1/150.

Higgins, D G. 2000. Amino acid-based phylogeny and alignment. pp 99-135 in *Analysis of Amino Acid Sequences* ed: Bork, P. Advances in Protein Chemistry. Vol: 54. Academic Press (San Diego).

de la Higuera, C; Casacuberta, F. 2000 Jan 6. Topology of strings: Median string is NP-complete. *Theoretical Computer Science* 230(1-2):39-48.

Hilser, V J; Thompson, E B. 2007 May 15. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proceedings of the National Academy of Sciences USA* 104(20):8311-8315. http://www.pnas.org/cgi/content/full/104/20/8311.

Hinds, D A; Levitt, M. 1996 Apr 26. From structure to sequence and back again. *Journal of Molecular Biology* 258(1):201-209.

Hoegger, P J; Kilaru, S; James, T Y; Thacker, J R; Kues, U. 2006 May. Phylogenetic comparison and classification of laccase and related multicopper oxidase protein sequences. *FEBS Journal* 273(10):2308-2326.

Holmes, I; Bruno, W J. 2001 Sep. Evolutionary HMMs: A Bayesian approach to multiple alignment. *Bioinformatics* 17(9):803-820. http://bioinformatics.oupjournals.org/cgi/content/abstract/17/9/803.

Hombrados, I; Rodewald, K; Neuzil, E; Braunitzer, G. 1983 Apr/May. Haemoglobins, LX. Primary structure of the major haemoglobin of the sea lamprey *Petromyzon marinus* (var. Garonne, Loire). *Biochimie* 65(4-5):247-257.

Huang, J-T; Wang, M-T. 2002 Jun 14. Secondary structural wobble: The limits of protein prediction accuracy. *Biochemical & Biophysical Research Communications* 294(3):621-625.

Hubbard, T J P; Ailey, B G; Brenner, S E; Murzin, A G; Chothia, C. 1999 Jan 1. SCOP: A structural classification of proteins database. *Nucleic Acids Research* 27(1):254-256. http://compbio.berkeley.edu/people/brenner/ http://nar.oupjournals.org/cgi/content/abstract/27/1/254.

Huelsenbeck, J P. 1995 Mar. Performance of phylogenetic methods in simulation. *Systematic Biology* 44(1):17-48.

Huelsenbeck, J P. 1997 Mar. Is the Felsenstein zone a fly trap? *Systematic Biology* 46(1):69-74.

Huelsenbeck, J P; Nielsen, R. 1999 Jan. Variation in the pattern of nucleotide substitution across sites. *Journal of Molecular Evolution* 48(1):86-93.

Huelsenbeck, J P; Ronquist, F. 2001 Aug 1. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8):754-755. http://bioinformatics.oupjournals.org/cgi/content/full/17/8/754.

Huelsenbeck, J P. 2002 May. Testing a covariotide model of DNA substitution. *Molecular Biology and Evolution* 19(5):698-707. http://mbe.oupjournals.org/cgi/content/full/19/5/698.

Huelsenbeck, J P; Bollback, J P; Levine, A M. 2002 Jan. Inferring the root of a phylogenetic tree. *Systematic Biology* 51(1):32-43.

Huelsenbeck, J P; Ronquist, F; van der Mark, P; Larget, B; Simon, D. 2006. MRBAYES, 3.1.2. Florida State University (Tallahassee, FL). http://mrbayes.csit.fsu.edu/.

Huelsenbeck, J P; Ronquist, F; van der Mark, P; Larget, B; Simon, D. 2007. MRBAYES, 3.2. Florida State University (Tallahassee, FL). http://mrbayes.csit.fsu.edu/.

Hurley, J H; Chen, R; Dean, A M. 1996 May 7. Determinants of cofactor specificity in isocitrate dehydrogenase: Structure of an engineered $NADP^+ \rightarrow NAD^+$ specificity-reversal mutant. *Biochemistry* 35(18):5670-5678.

Huson, D H; Nettles, S M; Warnow, T J. 1999 Fall/Winter. Disk-Covering: A fast-converging method for phylogenetic tree reconstruction. *Journal of Computational Biology* 6(3-4):369-386.

Itoh, T; Martin, W; Nei, M. 2002 Oct 1. Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proceedings of the National Academy of Sciences USA* 99(20):12944-12948. http://www.pnas.org/cgi/content/full/99/20/12944.

IUBMB, N C. 1992. *Enzyme Nomenclature*. Edition: 2007 (Online). Academic Press (San Diego). http://www.chem.qmul.ac.uk/iubmb/enzyme/.

Iyer, L M; Burroughs, A M; Aravind, L. 2006. The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains. *Genome Biology* 7(7):R60. http://genomebiology.com/2006/7/7/R60.

Jacoboni, I; Martelli, P L; Fariselli, P; Compiani, M; Casadio, R. 2000 Dec 1. Prediction of protein segments with the same amino acid sequence and different secondary structure: A benchmark for predictive methods. *PROTEINS: Structure, Function, and Genetics* 41(4):535-544.

Jaroszewski, L; Rychlewski, L; Godzik, A. 2000 Aug 1. Improving the quality of twilight-zone alignments. *Protein Science* 9(8):1487-1496. http://www.proteinscience.org/cgi/content/full/9/8/1487.

Jennings, A J; Edge, C M; Sternberg, M J E. 2001 Apr. An approach to improving multiple alignments of protein sequences via predicted secondary structure. *Protein Engineering* 14(4):227-231. http://peds.oxfordjournals.org/cgi/content/full/14/4/227.

Jermann, T M; Opitz, J G; Stackhouse, J; Benner, S A. 1995 Mar 2. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374(6517):57-59.

Jin, L; Nei, M. 1990 Jan. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular Biology and Evolution* 7(1):82-102. http://mbe.oupjournals.org/cgi/content/abstract/7/1/82.

John, B; Sali, A. 2003 Jul 15. Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Research* 31(14):3982-3992. http://nar.oupjournals.org/cgi/content/full/31/14/3982.

Johnson, M S; Overington, J P. 1993 Oct 20. A structural basis for sequence comparisons: An evaluation of scoring methodologies. *Journal of Molecular Biology* 233(4):716-738.

Johnson, S A S; Dubeau, L; White, R J; Johnson, D L. 2003 Sep/Oct. The TATA-Binding Protein as a regulator of cellular transformation. *Cell Cycle* 2(5):442-444. http://www.landesbioscience.com/journals/cc/abstract.php?id=493.

Jones, B E; Jennings, P A; Pierre, R A; Matthews, C R. 1994 Dec 27. Development of nonpolar surfaces in the folding of *Escherichia coli* dihydrofolate reductase detected by 1-anilinonaphthalene-8-sulfonate binding. *Biochemistry* 33(51):15250-15258.

Jones, M; Blaxter, M L. 2005 Apr 28. Evolutionary biology: Animal roots and shoots. *Nature* 434(7037):1076-1077.

Jorgensen, W L; Tirado-Rives, J. 1988 Mar 16. The OPLS potential functions for proteins: Energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society* 110(6):1657-1666.

Jowsey, I R; Thomson, A M; Flanagan, J U; Murdock, P R; Moore, G B T; Meyer, D J; Murphy, G J; Smith, S A; Hayes, J D. 2001 Nov 1. Mammalian class Sigma glutathione S-transferases: Catalytic properties and tissue-specific expression of human and rat GSH-dependent prostaglandin D2 synthases. *Biochemical Journal* 359(3):507-516. http://www.biochemj.org/bj/359/0507/bj3590507.htm.

Kabsch, W; Sander, C. 1983 Dec. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12):2577-2637. http://swift.cmbi.ru.nl/gv/dssp/.

Kahn, P C. 2006. Solvent-accessible surface areas of fully-extended residues. Personal communication. To: Smith, A W (New Brunswick, NJ). kahn@aesop.rutgers.edu.

Kahn, P C. 2007a. 2 layers of water needs at least 1.0 nm (ideally, 1.4 nm) of room on each side of the protein. Oral communication. To: Smith, A W (New Brunswick, NJ).

Kahn, P C. 2007b. 3 residues before/after should be enough (although 5 ideal). Oral communication. To: Smith, A W (New Brunswick, NJ).

Kahn, P C. 2007c Oral communication. 3.8 Ang. is reasonable for H-bond maximum distance. To: Smith, A W (New Brunswick, NJ).

Kahn, P C. 2007d. InsightII can't handle lots of water. Oral communication. To: Smith, A W (New Brunswick, NJ).

Kahn, P C. 2007e. Looking at pretty pictures doesn't do any good for evaluating models. Oral communication. To: Smith, A W (New Brunswick, NJ).

Kahn, P C. 2007f. Volume programs are limited to buried residues. Oral communication. To: Smith, A W (New Brunswick, NJ).

Kampfer, P. 2006. The Family Streptomycetaceae, Part I: Taxonomy. pp 538-604 in *Archaea; Bacteria: Firmicutes, Actinomycetes* ed: Dworkin, M; Falkow, S; Rosenberg, E; Schleifer, K-H; Stackebrandt, E. Edition: 3rd. The Prokaryotes. Vol: 3. Springer-Verlag (New York).

Karlin, S; Zhu, Z-Y; Baud, F. 1999 Oct 26. Atom density in protein structures. *Proceedings of the National Academy of Sciences USA* 96(22):12500-12505. http://www.pnas.org/cgi/content/full/96/22/12500.

Katoh, K; Misawa, K; Kuma, K-I; Miyata, T. 2002 Jul 15. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30(14):3059-3066. http://nar.oupjournals.org/cgi/content/full/30/14/3059.

Kawabata, T; Nishikawa, K. 2000 Oct 1. Protein structure comparison using the Markov transition model of evolution. *PROTEINS: Structure, Function, and Genetics* 41(1):108-122.

Kawase, T; Saito, A; Sato, T; Kanai, R; Fujii, T; Nikaidou, N; Miyashita, K; Watanabe, T. 2004 Feb 1. Distribution and phylogenetic analysis of Family 19 chitinases in Actinobacteria. *Applied and Environmental Microbiology* 70(2):1135-1144.

Keller, I; Benasasson, D; Nichols, R A. 2007 Feb 2. Transition-transversion bias is not universal: A counter-example from grasshopper pseudogenes. *PLoS Genetics* 3(2):e22. http://dx.doi.org/10.1371/journal.pgen.0030022.

Khalid, A. 2001. *Purification and characterization of aromatic alcohol:NADP + oxidoreductase from Escherichia coli; Cloning of the CAD gene and enzyme purification*. Mabel Smith Douglass Honors Thesis. Rutgers University (New Brunswick, NJ): Biochemistry and Microbiology, Douglass College.

Kim, B; Sahin, N; Minnikin, D; Zakrzewska-Czerwinska, J; Mordarski, M; Goodfellow, M. 1999 Jan 1. Classification of thermophilic streptomycetes, including the description of *Streptomyces thermoalcalitolerans* sp. nov. *International Journal of Systematic Bacteriology* 49(1):7-17. http://ijs.sgmjournals.org/cgi/content/abstract/49/1/7.

Kimsey, H H; Kaiser, D. 1992 Jan 15. The orotidine-5'-monophosphate decarboxylate gene of *Myxococcus xanthus*: Comparison to the OMP decarboxylase gene family. *Journal of Biological Chemistry* 267(2):819-824. http://www.jbc.org/cgi/content/abstract/267/2/819.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press (Cambridge). ISBN 0-521-23109-4.

Kirk, P; Gams, W; Stalpers, J; Stegehuis, G; Pennycook, S; Johnson, P; Cooper, J; Hawksworth, D L. 2007. Index Fungorum. Index Fungorum Partnership: CABI Bioscience, CBS (Centraalbureau voor Schimmelcultures), Landcare Research. http://www.indexfungorum.org.

Kirkpatrick, S; Gelatt, C D, Jr; Vecchi, M P. 1983 May 13. Optimization by simulated annealing. *Science* 220(4598):671-680. http://citeseer.ist.psu.edu/kirkpatrick83optimization.html http://www.cs.virginia.edu/cs432/documents/sa-1983.pdf http://en.wikipedia.org/wiki/Simulated_annealing.

Kishino, H; Hasegawa, M. 1989 Aug. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution* 29(2):170-179.

Kjer, K M. 1995 Sep. Use of ribosomal-RNA secondary structure in phylogenetic studies to identify homologous positions: An example of alignment and data presentation from the frogs. *Molecular Phylogenetics and Evolution* 4(3):314-330.

Kjer, K M. 1997 Dec. Conserved primary and secondary structural motifs of amphibian 12S rRNA, domain III. *Journal of Herpetology* 31(4):599-604.

Kjer, K M. 2007 Mar 6. Overparameterization. Email Communication. To: Smith, A W (New Brunswick, NJ).

Kleywegt, G J; Jones, T A. 1995 Jun 1. When freedom is given, liberties are taken. *Structure (Cambridge)* 3(6):535-540.

Kleywegt, G J; Brunger, A T. 1996 Aug 15. Checking your imagination: Applications of the free R value. *Structure with Folding & Design* 4(8):897-904.

Kloczkowski, A; Ting, T L; Jernigan, R L; Garnier, J. 2002 Nov 1. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *PROTEINS: Structure, Function, and Genetics* 49(2):154-166.

Knudsen, B; Miyamoto, M M. 2001 Dec 4. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proceedings of the National Academy of Sciences USA* 98(25):14512-14517. http://www.pnas.org/cgi/content/abstract/98/25/14512.

Koehl, P; Levitt, M. 2002 Jan 22. Improved recognition of native-like protein structures using a family of designed sequences. *Proceedings of the National Academy of Sciences USA* 99(2):691-696. http://www.pnas.org/cgi/content/full/99/2/691.

Kono, H; Saven, J G. 2001 Feb 23. Statistical theory for protein combinatorial libraries: Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *Journal of Molecular Biology* 306(3):607-628.

Koshi, J M; Goldstein, R A. 1995 Jul. Context-dependent optimal substitution matrices. *Protein Engineering* 8(7):641-645.

Koshi, J M; Mindell, D P; Goldstein, R A. 1997 Dec 12-13. Beyond mutation matrices: Physical-chemistry based evolutionary models. *Genome Informatics (GIW)* 8. Tokyo, Japan. Universal Academy Press. http://www.jsbi.org/journal/GI08.html http://www.umich.edu/~goldgrp/grouppubs.html http://giw.ims.u-tokyo.ac.jp/giw97/.

Koshi, J M; Goldstein, R A. 1998 Aug 15. Models of natural mutations including site heterogeneity. *PROTEINS: Structure, Function, and Genetics* 32(3):289-295. http://www.umich.edu/~goldgrp/grouppubs.html.

Koshi, J M; Mindell, D P; Goldstein, R A. 1999 Feb. Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes. *Molecular Biology and Evolution* 16(2):173-179. http://www.umich.edu/~goldgrp/grouppubs.html http://mbe.oupjournals.org/cgi/content/abstract/16/2/173.

Krahn, J M; Jackson, M R; DeRose, E F; Howell, E E; London, R E. 2007 Dec 25. Crystal structure of a Type II dihydrofolate reductase catalytic ternary complex. *Biochemistry* 46(51):14878-14888. http://pubs.acs.org/cgi-bin/article.cgi/bichaw/2007/46/i51/html/bi701532r.html.

Kraus, F; Jarecki, L; Miyamoto, M M; Tanhauser, S M; Laipis, P J. 1992 Jul. Mispairing and compensational changes during the evolution of mitochondrial ribosomal RNA. *Molecular Biology and Evolution* 9(4):770-774. http://mbe.oupjournals.org/cgi/content/abstract/9/4/770.

Kreitman, M; Comeron, J M. 1999 Dec. Coding sequence evolution. *Current Opinion in Genetics & Development* 9(6):637-641.

Krungkrai, J; Wutipraditkul, N; Prapunwattana, P; Krungkrai, S R; Rochanakij, S. 2001 Dec 15. A nonradioactive high-performance liquid chromatographic microassay for uridine 5'-monophosphate synthase, orotate phosphoribosyltransferase, and orotidine 5'-monophosphate decarboxylase. *Analytical Biochemistry* 299(2):162-168.

Kuhnert, P; Korczak, B M. 2006 Sep 1. Prediction of whole-genome DNA-DNA similarity, determination of G+C content and phylogenetic analysis within the family Pasteurellaceae by multilocus sequence analysis (MLSA). *Microbiology* 152(9):2537-2548.

Kullberg, M; Nilsson, M A; Arnason, U; Harley, E H; Janke, A. 2006 Aug. Housekeeping genes for phylogenetic analysis of eutherian relationships. *Molecular Biology and Evolution* 23(8):1493-1503. http://mbe.oxfordjournals.org/cgi/content/full/23/8/1493.

Kumar, S S C; Bansal, M. 1998a Jun 1. Dissecting alpha-helices: Position-specific analysis of alpha-helices in globular proteins. *PROTEINS: Structure, Function, and Genetics* 31(4):460-476.

Kumar, S S C; Bansal, M. 1998b Oct. Geometrical and sequence characteristics of alpha-helices in globular proteins. *Biophysical Journal* 75(4):1935-1944. http://www.biophysj.org/cgi/content/full/75/4/1935.

Kunsch, H R. 1989 Sep. The jacknife and the bootstrap for general stationary observations. *Annals of Statistics* 17(3):1217-1241.

Lake, J A. 1991 May. The order of sequence alignment can bias the selection of tree topology. *Molecular Biology and Evolution* 8(3):378-385. http://mbe.oupjournals.org/cgi/content/abstract/8/3/378.

Lake, J A. 1995 Oct 10. Calculating the probability of multitaxon evolutionary trees: Bootstrapper's gambit. *Proceedings of the National Academy of Sciences USA* 92(21):9662-9666. http://www.pnas.org/cgi/content/abstract/92/21/9662.

Landan, G; Graur, D. 2007 Jun 1. Heads or tails: A simple reliability check for multiple sequence alignments. *Molecular Biology and Evolution* 24(6):1380-1383.

Lanyon, S M. 1993 May. Phylogenetic frameworks: Toward a firmer foundation for the comparative approach. *Biological Journal of the Linnean Society* 49(1):45-61.

Lartillot, N; Brinkmann, H; Philippe, H. 2007 Feb 8. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BiomedCentral Evolutionary Biology* 7(Suppl 1):S4. http://www.biomedcentral.com/1471-2148/7/S1/S4.

Lathrop, R H. 1994 Sep. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Engineering* 7(9):1059-1068.

Lathrop, R H; Rogers, R G, Jr; Smith, T F; White, J V. 1998 Nov. A Bayes-optimal sequence-structure theory that unifies protein sequence-structure recognition and alignment. *Bulletin of Mathematical Biology* 60(6):1039-1071.

Lazar, G A; Desjarlais, J R; Handel, T M. 1997 Jun. *De novo* design of the hydrophobic core of ubiquitin. *Protein Science* 6(6):1167-1178.

Le Gall, T; Romero, P R; Cortese, M S; Uversky, V N; Dunker, A K. 2007 Feb. Intrinsic disorder in the Protein Data Bank. *Journal of Biomolecular Structure & Dynamics* 24(4):325-342.

Lee, B-K; Richards, F M. 1971 Feb 14. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology* 55(3):379-400. http://www.csb.yale.edu/userguides/datamanip/access/access_descrip.html.

Lee, J K; Houk, K N. 1997 May 9. A proficient enzyme revisited: The predicted mechanism for orotidine monophosphate decarboxylase. *Science* 276(5314):942-945.

Levin, J M; Pascarella, S; Argos, P; Garnier, J. 1993 Nov. Quantification of secondary structure prediction improvement using multiple alignments. *Protein Engineering* 6(8):849-854.

Levitt, M; Gerstein, M. 1998 May 26. A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of Sciences USA* 95(11):5913-5920. http://bioinfo.mbb.yale.edu/papers/ http://www.pnas.org/cgi/content/full/95/11/5913.

Lewis, W S; Cody, V; Galitsky, N; Luft, J R; Pangborn, W; Chunduru, S K; Spencer, H T; Appleman, J R; Blakley, R L. 1995 Mar 10. Methotrexate-resistant variants of human dihydrofolate reductase with substitutions of leucine 22: Kinetics, crystallography, and potential as selectable markers. *Journal of Biological Chemistry* 270(10):5057-5064. http://www.jbc.org/content/vol270/issue10/.

Lin, J. 1991 Jan. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37(1):145-151.

Lindahl, E; Hess, B; van der Spoel, D. 2001. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling* 7:306-317. http://folding.bmc.uu.se/ http://www.gromacs.org.

Lindahl, E; van der Spoel, D; Hess, B; Groenhof, G; Kutzner, C. 2007. GROMACS, 3.3.2. http://www.gromacs.org.

Lio, P; Goldman, N. 1998 Dec. Models of molecular evolution and phylogeny. *Genome Research* 8(12):1233-1244. http://www.genome.org/cgi/content/full/8/12/1233.

Lio, P; Goldman, N; Thorne, J L; Jones, D T. 1998 Sep. PASSML: Combining evolutionary inference and protein secondary structure prediction. *Bioinformatics* 14(8):726-733. http://bioinformatics.oupjournals.org/cgi/content/abstract/14/8/726.

Lipke, P N; Chen, M-H; de Nobel, H; Kurjan, J; Kahn, P C. 1995 Oct. Homology modeling of an immunoglobulin-like domain in the *Saccharomyces cerevisiae* adhesion protein alpha-agglutinin. *Protein Science* 4(10):2168-2178. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2142996 http://www.proteinscience.org/cgi/content/abstract/4/10/2168.

Liu, X Z; Zhang, L M; Guan, S; Zheng, W M. 2003 May. Distances and classification of amino acids for different protein secondary structures. *Physics Reviews E* 67(5 Pt 1):051927. http://en.wikipedia.org/wiki/Kullback-Leibler_divergence.

Liu, Z; Shi, Y; Zhang, Y; Zhou, Z; Lu, Z; Li, W; Huang, Y; Rodriguez, C; Goodfellow, M. 2005 Jul 1. Classification of *Streptomyces griseus* (Krainsky 1914) Waksman and Henrici 1948 and related species and the transfer of '*Microstreptospora cinerea*' to the genus *Streptomyces* as *Streptomyces yanii sp. nov. International Journal of Systematic and Evolutionary Microbiology* 55(4):1605-1610.

Lo Conte, L; Ailey, B G; Hubbard, T J P; Brenner, S E; Murzin, A G; Chothia, C. 2000 Jan 1. SCOP: A structural classification of proteins database. *Nucleic Acids Research* 28(1):257-259. http://nar.oupjournals.org/content/vol28/issue1/ http://compbio.berkeley.edu/people/brenner/.

Looger, L L; Hellinga, H W. 2001 Mar 16. Generalized Dead-End Elimination algorithms make large-scale protein side-chain structure prediction tractable: Implications for protein design and structural genomics. *Journal of Molecular Biology* 307(1):429-445.

Lopez-Ortiz, A. 2000 May 27. Comp.Theory FAQ: Frequently Asked Questions about theoretical computer science. http://www.cs.uwaterloo.ca/~alopez-o/comp-faq/faq.html.

Lopez, P; Forterre, P; Philippe, H. 1999 Oct. The root of the tree of life in the light of the covarion model. *Journal of Molecular Evolution* 49(4):496-508.

Lovell, S C; Word, J M; Richardson, J S; Richardson, D C. 1999 Jan 19. Asparagine and glutamine rotamers: B-factor cutoff and correction of amide flips yield distinct clustering. *Proceedings of the National Academy of Sciences USA* 96(2):400-405. http://www.pnas.org/cgi/content/abstract/96/2/400.

Lovell, S C; Word, J M; Richardson, J S; Richardson, D C. 2000 Aug 15. The penultimate rotamer library. *PROTEINS: Structure, Function, and Genetics* 40(3):389-408.

Lovell, S C; Davis, I W; Arendall, W B, III; de Bakker, P I W; Word, J M; Prisant, M G; Richardson, J S; Richardson, D C. 2003 Feb 15. Structure validation by C_{alpha} geometry: Phi, Psi and $C_{ß}$ deviation. *PROTEINS: Structure, Function, and Genetics* 50(3):437-450.

Lundstrom, J; Rychlewski, L; Bujnicki, J; Elofsson, A. 2001 Nov. Pcons: A neural-network-based consensus predictor that improves fold recognition. *Protein Science* 10(11):2354-2362. http://www.proteinscience.org/cgi/content/full/10/11/2354.

Luzzati, V. 1952 Nov. Traitement statistique des erreurs dans la determination des structures cristallines. *Acta Crystallographica* 5(6):802-810. http://dx.doi.org/10.1107/S0365110X52002161.

Luzzati, V. 1953 Feb. Resolution d'un structure cristalline lorsque les positions d'une partie des atoms sont connues: Traitement statistique. *Acta Crystallographica* 6(2):142-152. http://dx.doi.org/10.1107/S0365110X53000508.

Ma, L; Imamichi, H; Sukura, A; Kovacs, J A. 2001 Nov 15. Genetic divergence of the dihydrofolate reductase and Dihydropteroate Synthase genes in *Pneumocystis carinii* from 7 different host species. *Journal of Infectious Diseases* 184(10):1358-1362.

Maddison, D R; Swofford, D L; Maddison, W P. 1997 Dec. NEXUS: An extensible file format for systematic information. *Systematic Biology* 46(4):590-621.

Malthus, T R. 1798. *An Essay on the Principle of Population, as it affects the Future Improvement of Society with remarks on the Speculations of Mr. Godwin, M. Condorcet, and Other Writers*. Edition: 1st. J. Johnson (London). http://www.econlib.org/library/Malthus/malPop.html.

Margulis, L. 1996 Feb 6. Archaeal-eubacterial mergers in the origin of Eukarya: Phylogenetic classification of life. *Proceedings of the National Academy of Sciences USA* 93(3):1071-1076. http://www.pnas.org/cgi/content/abstract/93/3/1071.

Marshall, C R; Raff, E C; Raff, R A. 1994 Dec 6. Dollo's Law and the death and resurrection of genes. *Proceedings of the National Academy of Sciences USA* 91(25):12283-12287. http://www.pnas.org/cgi/content/abstract/91/25/12283.

Marti-Renom, M A; Stuart, A C; Fiser, A; Sanchez, R; Melo, F; Sali, A. 2000. Comparative protein structure modeling of genes and genomes. pp 291-325 ed: Stroud, R M; Olson, W K; Sheetz, M P. Annual Review of Biophysics and Biomolecular Structure. Vol: 29. Annual Reviews (Palo Alto, CA).

Massey, S E; Moura, G; Beltrao, P; Almeida, R; Garey, J R; Tuite, M F; Santos, M A S. 2003 Apr 1. Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in *Candida spp. Genome Research* 13(4):544-557. http://www.genome.org/cgi/content/full/13/4/544.

Matheny, P B; Wang, Z; Binder, M; Curtis, J M; Lim, Y W; Henrik Nilsson, R; Hughes, K W; Hofstetter, V; Ammirati, J F; Schoch, C L; Langer, E; Langer, G; McLaughlin, D J; Wilson, A W; Froslev, T; Ge, Z-W; Kerrigan, R W; Slot, J C; Yang, Z-L; Baroni, T J; Fischer, M; Hosaka, K; Matsuura, K; Seidl, M T; Vauras, J; Hibbett, D S. 2007 May. Contributions of rpb2 and tef1 to the phylogeny of mushrooms and allies (Basidiomycota, Fungi). *Molecular Phylogenetics and Evolution* 43(2):430-451.

McLachlan, A D. 1984 Jul. How alike are the shapes of two random chains? *Biopolymers* 23(7):1325-1331.

McTique, M A; Davies, J F, 2nd; Kaufman, B T; Kraut, J. 1993 Jul 13. Crystal structures of chicken liver dihydrofolate reductase: Binary thioNADP+ and ternary thioNADP+.biopterin complexes. *Biochemistry* 32(27):6855-6862.

Merali, S; Frevert, U; Williams, J H; Chin, K; Bryan, R; Clarkson, A B, Jr. 1999 Mar 2. Continuous axenic cultivation of *Pneumocystis carinii*. *Proceedings of the National Academy of Sciences USA* 96(5):2402-2407. http://www.pnas.org/cgi/content/full/96/5/2402.

Metropolis, N; Rosenbluth, A W; Rosenbluth, M N; Teller, A H; Teller, E. 1953 Jun. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21(6):1087-1092. http://en.wikipedia.org/wiki/Metropolis-Hastings_algorithm.

Metsa-Ketela, M; Halo, L; Munukka, E; Hakala, J; Mantsala, P; Ylihonko, K. 2002 Sep 1. Molecular evolution of aromatic polyketides and comparative sequence analysis of Polyketide Ketosynthase and 16S Ribosomal DNA genes from various *Streptomyces* species. *Applied and Environmental Microbiology* 68(9):4472-4479.

Meyer, S; von Haeseler, A. 2003 Feb 1. Identifying site-specific substitution rates. *Molecular Biology and Evolution* 20(2):182-189. http://mbe.oupjournals.org/cgi/content/full/20/2/182.

Mi, S; Lee, X; Li, X-p; Veldman, G M; Finnerty, H; Racie, L; LaVallie, E; Tang, X-Y; Edouard, P; Howes, S; Keith, J C, Jr; McCoy, J M. 2000 Feb 17. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* 403(6771):785-789.

Miller, B G; Smiley, J A; Short, S A; Wolfenden, R. 1999 Aug 20. Activity of yeast orotidine-5'-phosphate decarboxylase in the absence of metals. *Journal of Biological Chemistry* 274(34):23841-23843. http://www.jbc.org/content/vol274/issue34/.

Miller, B G; Hassell, A M; Milburn, M V; Short, S A. 2000a Apr. Crystallization of native and selenomethionyl yeast orotidine 5'-phosphate decarboxylase. *Acta Crystallographica Section D* 56(4):472-474.

Miller, B G; Hassell, A M; Wolfenden, R; Milburn, M V; Short, S A. 2000b Feb 29. Anatomy of a proficient enzyme: The structure of orotidine 5'-monophosphate decarboxylase in the presence and absence of a potential transition state analog. *Proceedings of the National Academy of Sciences USA* 97(5):2011-2016. http://www.pnas.org/cgi/content/full/97/5/2011.

Miller, B G; Snider, M J; Short, S A; Wolfenden, R. 2000c Jul 18. Contribution of enzyme-phosphoribosyl contacts to catalysis by orotidine 5'-phosphate decarboxylase. *Biochemistry* 39(28):8113-8118.

Mirny, L A; Shakhnovich, E I. 1998 Oct 23. Protein structure prediction by threading: Why it works and why it does not. *Journal of Molecular Biology* 283(2):507-526.

Mitchison, G J; Durbin, R M. 1995 Dec. Tree-based maximal likelihood substitution matrices and hidden Markov models. *Journal of Molecular Evolution* 41(6):1139-1151.

Mitchison, G J. 1999 Jul. A probabilistic treatment of phylogeny and sequence alignment. *Journal of Molecular Evolution* 49(1):11-22.

Mito, M; Chong, K T; Miyazaki, G; Adachi, S-i; Park, S-Y; Tame, J R H; Morimoto, H. 2002 Jun 14. Crystal structures of deoxy- and carbonmonoxyhemoglobin F1 from the hagfish *Epatretus burgeri*. *Journal of Biological Chemistry* 277(24):21898-21905. http://www.jbc.org/cgi/content/full/277/24/21898.

Miyazaki, J; Nakaya, S; Suzuki, T; Tamakoshi, M; Tanabe, Y; Oshima, T; Yamagishi, A. 2001 May. Ancestral residues stabilizing 3-isopropylmalate dehydrogenase of an extreme thermophile: Experimental evidence supporting the thermophilic common ancestor hypothesis. *Journal of Biochemistry* 129(5):777-782.

Mizuguchi, K; Deane, C M; Blundell, T L; Overington, J P. 1998 Nov. HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Science* 7(11):2469-2471. http://www.proteinscience.org/cgi/content/abstract/7/11/2469.

Mizuguchi, K; Blundell, T L. 2000 Dec. Analysis of conservation and substitutions of secondary structure elements within protein superfamilies. *Bioinformatics* 16(12):1111-1119. http://bioinformatics.oupjournals.org/cgi/content/abstract/16/12/1111.

Moran, N A. 1996 Apr 2. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences USA* 93(7):2873-2878. http://www.pnas.org/cgi/content/abstract/93/7/2873.

Moreira, D; Lopez-Garcia, P; Vickerman, K. 2004 Sep. An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: Proposal for a new classification of the class Kinetoplastea. *International Journal of Systematic and Evolutionary Microbiology* 54(5):1861-1875. http://ijs.sgmjournals.org/cgi/content/full/54/5/1861.

Moret, B M E; Roshan, U; Warnow, T J; Williams, T L. 2003. Performance of supertree methods on various dataset decompositions. in *Phylogenetic Supertrees* ed: Bininda-Emonds, O R P. Kluwer. http://www.cs.unm.edu/~moret/papers.html.

Morrison, D A; Ellis, J T. 1997 Apr 1. Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of apicomplexa. *Molecular Biology and Evolution* 14(4):428-441. http://mbe.oupjournals.org/cgi/content/abstract/14/4/428.

Mosimann, S; Meleshko, R; James, M N G. 1995 Nov. A critical assessment of comparative molecular modeling of tertiary structures of proteins. *PROTEINS: Structure, Function, and Genetics* 23(3):301-317.

Mouchacca, J. 2000a Oct 22. Thermophilic fungi and applied research: A synopsis of name changes and synonymies. *World Journal of Microbiology and Biotechnology* 16(8):881-888.

Mouchacca, J. 2000b Sep 22. Thermotolerant fungi erroneously reported in applied research work as possessing thermophilic attributes. *World Journal of Microbiology and Biotechnology* 16(8):869-880.

Mulder, N J; Apweiler, R; Attwood, T K; Bairoch, A; Bateman, A; Binns, D; Bork, P; Buillard, V; Cerutti, L; Copley, R; Courcelle, E; Das, U; Daugherty, L; Dibley, M; Finn, R; Fleischmann, W; Gough, J; Haft, D; Hulo, N; Hunter, S; Kahn, D; Kanapin, A; Kejariwal, A; Labarga, A; Langendijk-Genevaux, P S; Lonsdale, D; Lopez, R; Letunic, I; Madera, M; Maslen, J; McAnulla, C; McDowall, J; Mistry, J; Mitchell, A; Nikolskaya, A N; Orchard, S; Orengo, C; Petryszak, R; Selengut, J D; Sigrist, C J A; Thomas, P D; Valentin, F; Wilson, D; Wu, C H; Yeats, C. 2007 Jan 12. New

developments in the InterPro database. *Nucleic Acids Research* 35(Suppl 1):D224-D228. http://nar.oxfordjournals.org/cgi/content/abstract/35/suppl_1/D224.

Mullan, L J; Bleasby, A. 2002 Mar. Short EMBOSS user guide: European Molecular Biology Open Software Suite. *Briefings in Bioinformatics* 3(1):92-94. http://bib.oxfordjournals.org/cgi/reprint/3/1/92.

Murzin, A G; Lo Conte, L; Ailey, B G; Brenner, S E; Hubbard, T J P; Chothia, C. 2000 Jul 1. SCOP: Structural classification of proteins, 1.53 release. http://scop.mrc-lmb.cam.ac.uk/scop/ http://scop.berkeley.edu/.

Muse, S V; Gaut, B S. 1994 Sep. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* 11(5):715-724. http://mbe.oupjournals.org/cgi/content/abstract/11/5/715.

Naor, D; Fischer, D; Jernigan, R L; Wolfson, H J; Nussinov, R. 1996 Mar 16. Amino acid pair interchanges at spatially conserved locations. *Journal of Molecular Biology* 256(5):924-938. http://citeseer.ist.psu.edu/naor96amino.html.

Neher, E. 1994 Jan 4. How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences USA* 91(1):98-102. http://www.pnas.org/cgi/content/abstract/91/1/98.

Nei, M; Zhang, J; Yokoyama, S. 1997 Jun. Color vision of ancestral organisms of higher primates. *Molecular Biology and Evolution* 14(6):611-618. http://mbe.oupjournals.org/cgi/content/abstract/14/6/611.

Nei, M; Kumar, S S C. 2000a. Chapter 2: Evolutionary change of amino acid sequences. pp 17-32 in *Molecular Evolution and Phylogenetics*. Oxford University Press (New York).

Nei, M; Kumar, S S C. 2000b. Chapter 5: Phylogenetic trees. pp 73-85 in *Molecular Evolution and Phylogenetics*. Oxford University Press (New York).

Nelson, G. 1983. Reticulation in cladograms. pp 105-111 in *Proceedings of the Second Meeting of the Willi Hennig Society* ed: Platnick, N I; Funk, V A. Advances in Cladistics. Vol: 2. Columbia University Press (New York).

Nicodeme, P. 2001 Jun 1. Fast approximate motif statistics. *Journal of Computational Biology* 8(3):235-248.

O'hUigin, C; Satta, Y; Takahata, N; Klein, J. 2002 Sep 1. Contribution of homoplasy and of ancestral polymorphism to the evolution of genes in anthropoid primates. *Molecular Biology and Evolution* 19(9):1501-1513. http://mbe.oupjournals.org/cgi/content/full/19/9/1501.

Ohmstede, C A; Langdon, S D; Chae, C B; Jones, M E. 1986 Mar 25. Expression and sequence analysis of a cDNA encoding the orotidine-5'-monophosphate decarboxylase domain from *Ehrlich ascites* uridylate synthase. *Journal of Biological Chemistry* 261(9):4276-4282. http://www.jbc.org/cgi/content/abstract/261/9/4276.

Olivella, M; Deupi, X; Govaerts, C; Pardo, L. 2002 Jun. Influence of the environment in the conformation of alpha-helices studied by protein database search and molecular dynamics simulations. *Biophysical Journal* 82(6):3207-3213. http://www.biophysj.org/cgi/content/full/82/6/3207.

Oostenbrink, C; Villa, A; Mark, A E; van Gunsteren, W F. 2004 Oct. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *Journal of Computational Chemistry* 25(13):1656-1676.

Oostenbrink, C; Soares, T A; van der Veqt, N F A; van Gunsteren, W F. 2005 Jun. Validation of the 53A6 GROMOS force field. *European Biophysics Journal* 34(4):273-284.

Ortiz, A R; Kolinski, A; Rotkiewicz, P; Ilkowski, B; Skolnick, J. 1999. *Ab initio* folding of proteins using restraints derived from evolutionary information. *PROTEINS: Structure, Function, and Genetics* Suppl 3:177-185.

Ortlund, E A; Bridgham, J T; Redinbo, M R; Thornton, J W. 2007 Sep 14. Crystal structure of an ancient protein: Evolution by conformational epistasis. *Science* 317(5844):1544-1548.

Otto, S P; Cummings, M P; Wakeley, J. 1996. Inferring phylogenies from DNA sequence data: The effects of sampling. pp 103-115 in *New Uses for New Phylogenies* ed: Harvey, P H; Leigh-Brown, A J; Maynard-Smith, J; Nee, S. Oxford University Press (Oxford).

Overington, J P; Johnson, M S; Sali, A; Blundell, T L. 1990 Aug 22. Tertiary structural constraints on protein evolutionary diversity. *Proceedings of the Royal Society of London B* 241(1301):132-145.

Overington, J P; Donnelly, D; Johnson, M S; Sali, A; Blundell, T L. 1992 Feb. Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Science* 1(2):216-226. http://www.proteinscience.org/cgi/content/abstract/1/2/216.

Pace, C N; Scholtz, J M. 1998 Jul 1. A helix propensity scale based on experimental studies of peptides and proteins. *Biophysical Journal* 75(1):422-427. http://www.biophysj.org/cgi/content/full/75/1/422.

Page, R D M. 1993 Mar. Genes, organisms, and areas: The problem of multiple lineages. *Systematic Biology* 42(1):77-84.

Palleroni, N J. 2003 Jan 1. Prokaryote taxonomy of the 20th century and the impact of studies on the genus *Pseudomonas*: A personal view. *Microbiology* 149(1):1-7. http://mic.sgmjournals.org/cgi/content/abstract/149/1/1.

Parise, A P. 2005 Jul 1. Family 18 glycosyl hydrolases (chitinases). Oral Communication. To: Smith, A W (New Brunswick, NJ).

Pascarella, S; Milpetz, F; Argos, P. 1996 Mar. A databank (3D_ali) collecting related protein sequences and structures. *Protein Engineering* 9(3):249-251. http://peds.oxfordjournals.org/cgi/reprint/9/3/249.

Peng, C-K; Buldyrev, S V; Goldberger, A L; Havlin, S; Sciortino, F; Simons, M; Stanley, H E. 1992 Mar 12. Long-range correlations in nucleotide sequences. *Nature* 356(6365):168-170.

Penny, D; Hendy, M D. 1985. Testing methods of evolutionary tree construction. *Cladistics* 1:266-278.

Penq, K; Vucetic, S; Radivojac, P; Brown, C J; Dunker, A K; Obradovic, Z. 2005 Feb. Optimizing long intrinsic disorder predictors with protein evolutionary information. *Journal of Bioinformatics and Computational Biology* 3(1):35-60.

Penq, K; Radivojac, P; Vucetric, S; Dunker, A K; Obradovic, Z. 2006 Apr 17. Length-dependent prediction of protein intrinsic disorder. *BioMedCentral Bioinformatics* 7:208. http://www.biomedcentral.com/1471-2105/7/208.

Philippe, H; Laurent, J. 1998 Dec. How good are deep phylogenetic trees? *Current Opinion in Genetics & Development* 8(6):616-623.

Philippe, H; Lartillot, N; Brinkmann, H. 2005 May. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Molecular Biology and Evolution* 22(5):1246-1253. http://mbe.oxfordjournals.org/cgi/content/full/22/5/1246.

Piaggio-Talice, R; Piaggio, R. 2003. Quartet Suite: Supertrees by Quartets, 1.0. Iowa State University (Ames, Iowa). http://genome.cs.iastate.edu/CBL/download/.

Piaggio-Talice, R; Burleigh, J G; Eulenstein, O. 2004. Quartet supertrees. pp 173-191 in *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* ed: Bininda-Emonds, O R P. Kluwer (Dordrecht, Netherlands). http://www.cs.iastate.edu/~rpiaggio/.

Pollock, D D; Taylor, W R; Goldman, N. 1999 Mar. Coevolving protein residues: Maximum likelihood identification and relationship to structure. *Journal of Molecular Biology* 287(1):187-198.

Pollock, D D; Bruno, W J. 2000 Dec. Assessing an unknown evolutionary process: Effect of increasing site-specific knowledge through taxon addition. *Molecular Biology and Evolution* 17(12):1854-1858. http://mbe.oxfordjournals.org/cgi/content/full/17/12/1854.

Ponder, J W; Richards, F M. 1987a. Internal packing and protein structural classes. *Evolution of Catalytic Function* 52:421-428. Cold Spring Harbor Symposia on Quantitative Biology. Cold Spring Harbor Laboratory.

Ponder, J W; Richards, F M. 1987b Feb 20. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology* 193(4):775-793.

Porter, D J; Short, S A. 2000 Sep 26. Yeast orotidine-5'-phosphate decarboxylase: Steady-state and pre-steady-state analysis of the kinetic mechanism of substrate decarboxylation. *Biochemistry* 39(38):11788-11800.

Preisner, R; Goede, A; Michalski, E; Frommel, C. 1997 Sep 8. Inverse sequence similarity in proteins and its relation to the three-dimensional fold. *FEBS Letters* 414(2):425-429.

Prieto, J; Serrano, L. 1997 Nov 28. C-capping and helix stability: The pro C-capping motif. *Journal of Molecular Biology* 274(2):276-288.

Przybylski, D; Rost, B. 2002 Feb 1. Alignments grow, secondary structure prediction improves. *PROTEINS: Structure, Function, and Genetics* 46(2):197-205.

Pupko, T; Huchon, D; Cao, Y; Okada, N; Hasegawa, M. 2002 Dec. Combining multiple data sets in a likelihood analysis: Which models are the best? *Molecular Biology and Evolution* 19(12):2294-2307. http://mbe.oupjournals.org/cgi/content/full/19/12/2294.

Radford, A. 1993 Apr. A fungal phylogeny based upon orotidine 5'-monophosphate decarboxylase. *Journal of Molecular Evolution* 36(4):389-395.

Radivojac, P; Obradovic, Z; Smith, D K; Zhu, G; Vucetic, S; Brown, C J; Lawson, J D; Dunker, A K. 2004 Jan. Protein flexibility and intrinsic disorder. *Protein Science* 13(1):71-80. http://www.proteinscience.org/cgi/content/full/13/1/71.

Ramachandra, M; Crawford, D L; Pometto, A L, III. 1987 Dec 1. Extracellular enzyme activities during lignocellulose degradation by *Streptomyces* spp.: A comparative study of wild-type and genetically manipulated strains. *Applied and Environmental Microbiology* 53(12):2754-2760. http://aem.asm.org/cgi/content/abstract/53/12/2754.

Rambaut, A. 2000 Apr 1. Estimating the rate of evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* 16(4):395-399. http://bioinformatics.oupjournals.org/cgi/content/abstract/16/4/395.

Ranwez, V; Gascuel, O. 2001 Jun. Quartet-based phylogenetic inference: Improvements and limits. *Molecular Biology and Evolution* 18(6):1103-1116. http://mbe.oupjournals.org/cgi/content/full/18/6/1103.

Reardon, D; Farber, G K. 1995 Apr. Protein motifs 4: The structure and evolution of alpha/beta barrel proteins. *FASEB Journal* 9(7):497-503.

Redfield, R; Findlay, W; Bosse, J; Kroll, J S; Cameron, A; Nash, J. 2006. Evolution of competence and DNA uptake specificity in the Pasteurellaceae. *BiomedCentral Evolutionary Biology* 6(1):82. http://www.biomedcentral.com/1471-2148/6/82.

Remington, S J; Matthews, B W. 1980 Jun 15. A systematic approach to the comparison of protein structures. *Journal of Molecular Biology* 140(1):77-99.

Rhodes, G. 2000. *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*. Edition: 2nd. Academic Press (New York). http://www.usm.maine.edu/~rhodes/CMCC/index.html.

Rice, D W; Eisenberg, D S. 1997 Apr 11. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *Journal of Molecular Biology* 267(4):1026-1038.

Rice, P; Longden, I; Bleasby, A. 2000 Jun. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16(6):276-277. http://emboss.sourceforge.net/.

Richards, F M. 1974 Jan 5. The interpretation of protein structures; Total volume, group volume distributions, and packing density. *Journal of Molecular Biology* 82(1):1-14. http://www.csb.yale.edu/userguides/datamanip/volume/volume_descrip.html.

Richardson, D C; Richardson, J S. 1992 Jan. The kinemage: A tool for scientific communication. *Protein Science* 1(1):3-9. http://www.proteinscience.org/cgi/content/abstract/1/1/3.

Richardson, D C; Richardson, J S. 2001 Apr 03. The Richardsons' 3-D protein structure homepage. Duke University. http://kinemage.biochem.duke.edu/.

Richardson, D C. 2007. KiNG, 2.13. Duke University. http://kinemage.biochem.duke.edu/software/king.php.

Richardson, J S. 1981, 2004-2006. The anatomy and taxonomy of protein structure. pp 246-253 ed: Anfinsen, C B; Edsall, J T; Richards, F M. Advances in Protein Chemistry. Vol: 34. Academic Press (New York). http://kinemage.biochem.duke.edu/teaching/anatax/index.html.

Richardson, J S; Richardson, D C. 2002 Mar 5. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proceedings of the National Academy of Sciences USA* 99(5):2754-2759. http://www.pnas.org/cgi/content/full/99/5/2754.

Rimet, O; Chauvet, M; Bourdeaux, M; Briand, C. 1987 Sep. A novel fluorometric assay for quantitative analysis of dihydrofolate reductase activity in biological samples. *Journal of Biochemical and Biophysical Methods* 14(6):335-342.

Rimet, O; Chauvet, M; Bourdeaux, M; Briand, C; Sastre, B. 1990 Jul. Activity measurements in human tissues of the methotrexate molecular target: A novel fluorometric assay. *Cancer Biochemistry and Biophysics* 11(3):239-245.

Rimet, O; Chauvet, M; Sarrazin, M; Bourdeaux, M. 1991 Feb 15. Conformational change induced by coenzyme binding to bovine liver dihydrofolate reductase: A spectrofluorimetric study. *Biochimica et Biophysica Acta* 1076(3):435-438.

Rishavy, M A; Cleland, W W. 2000 Apr 25. Determination of the mechanism of orotidine 5'-monophosphate decarboxylase by isotope effects. *Biochemistry* 39(16):4569-4574.

Robinson, D M; Jones, D T; Kishino, H; Goldman, N; Thorne, J L. 2003 Oct 1. Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution* 20(10):1692-1704. http://mbe.oupjournals.org/cgi/content/full/20/10/1692.

Robinson, M; Gouy, M; Gautier, C; Mouchiroud, D. 1998 Sep 1. Sensitivity of the relative-rate test to taxonomic sampling. *Molecular Biology and Evolution* 15(9):1091-1091. http://mbe.oupjournals.org/cgi/content/abstract/15/9/1091.

Rodriguez-Trelles, F; Tarrio, R; Ayala, F J. 2001 Sep 25. Erratic overdispersion of three molecular clocks: GPDH, SOD, and XDH. *Proceedings of the National Academy of Sciences USA* 98(20):11405-11410. http://www.pnas.org/cgi/content/abstract/98/20/11405.

Romanelli, R A; Houston, C W; Barnett, S M. 1975 Aug. Studies on thermophilic cellulolytic fungi. *Applied Microbiology* 30(2):276-281. http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=16350026.

Romero, P R; Zaidi, S; Fang, Y Y; Uversky, V N; Radivojac, P; Oldfield, C J; Cortese, M S; Sickmeier, M; LeGall, T; Obradovic, Z; Dunker, A K. 2006 May 30. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proceedings of the National Academy of Sciences USA* 103(22):8390-8395. http://www.pnas.org/cgi/content/full/103/22/8390.

Ronquist, F; Huelsenbeck, J P. 2003 12 Aug. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572-1574. http://bioinformatics.oxfordjournals.org/cgi/content/full/19/12/1572.

Ronquist, F. 2004 Sep 2004. Bayesian inference of character evolution. *Trends in Ecology and Evolution* 19(9):475-481.

Ronquist, F. 2005 May. Bayesian phylogenetic inference introduction. Florida State University. http://people.scs.fsu.edu/~ronquist/mrbayes/ http://www.csit.fsu.edu/~ronquist/mrbayes/BayesianInference_I.ppt.

Rose, G D; Creamer, T P. 1994 May. Protein folding: Predicting predicting. *PROTEINS: Structure, Function, and Genetics* 19(1):1-3.

Roshan, U; Moret, B M E; Williams, T L; Warnow, T J. 2004a. Rec-I-DCM3: A fast algorithmic technique for reconstructing large phylogenetic trees. *IEEE Computational Systems Bioinformatics Conference*. Proceedings of the IEEE CSB Conference. IEEE Press. http://www.cs.njit.edu/usman/phylogenetics/csb04.pdf.

Roshan, U; Warnow, T J; Moret, B M E; Williams, T L. 2004b. REC-I-DCM3, 1.0. New Jersey Institute of Technology (Newark, NJ). http://www.cs.njit.edu/~usman/RecIDCM3.html.

Rossmann, M G; Moras, D; Olsen, K W. 1974 Jul 19. Chemical and biological evolution of a nucleotide-binding protein. *Nature* 250(463):194-199.

Rost, B; Sander, C. 1996. Bridging the protein sequence-structure gap by structure predictions. pp 113-136 ed: Stroud, R M; Hubbell, W L; Olson, W K; Sheetz, M P. Annual Review of Biophysics and Biomolecular Structure. Vol: 25. Annual Reviews (Palo Alto, CA).

Rost, B. 1999 Feb. Twilight zone of protein sequence alignments. *Protein Engineering* 12(2):85-94. http://www.columbia.edu/~rost/Papers/1999_twilight/paper.html.

Roy, M S; Geffen, E; Smith, E; Ostrander, E A; Wayne, R K. 1994 Jul. Patterns of differentiation and hybridization in North American wolflike canids, revealed by analysis of microsatellite loci. *Molecular Biology and Evolution* 11(4):553-570. http://mbe.oxfordjournals.org/cgi/content/abstract/11/4/553.

Ruiz-Trillo, I; Riutmort, M; Littlewood, D T J; Herniou, E A; Baquna, J. 1999 Mar 19. Acoel flatworms: Earliest extant bilaterian Metazoans, not members of Platyhelminthes. *Science* 283(5409):1919-1923. http://www.sciencemag.org/cgi/content/full/283/5409/1919.

Russell, R B; Barton, G J. 1993 Dec 20. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *Journal of Molecular Biology* 234(4):951-957.

Russell, R B; Saqi, M A S; Bates, P A; Sayle, R A; Sternberg, M J E. 1998 Jan. Recognition of analogous and homologous protein folds: Assessment of prediction success and associated alignment accuracy using empirical substitution matrices. *Protein Engineering* 11(1):1-9.

Russo, C A M; Takezaki, N; Nei, M. 1996 Mar. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Molecular Biology and Evolution* 13(3):525-536. http://mbe.oupjournals.org/cgi/content/abstract/13/3/525.

Sakai, Y; Kazarimoto, T; Tani, Y. 1991 Dec. Transformation system for an asporogenous methylotrophic yeast, *Candida boidinii*: Cloning of the orotidine-5'-phosphate decarboxylase gene (URA3), isolation of uracil auxotrophic mutants, and use of the mutants for integrative transformation. *Journal of Bacteriology* 173(23):7458-7463. http://jb.asm.org/cgi/content/abstract/173/23/7458 http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=1938943.

Salamov, A A; Solovyev, V V. 1995 Mar 17. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *Journal of Molecular Biology* 247(1):11-15.

Sali, A; Blundell, T L. 1993 Dec 5. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* 234(3):779-815. http://www.salilab.org/publications/ref/ http://www.salilab.org/publications/papers/model-93/.

Sali, A; Overington, J P. 1994 Sep. Derivation of rules for comparative protein modeling from a database of protein structural alignments. *Protein Science* 3(9):1582-1596. http://www.proteinscience.org/cgi/content/abstract/3/9/1582.

Sali, A. 1995 Sep. Protein modeling by satisfaction of spatial restraints. *Molecular Medicine Today* 1(6):270-277. http://www.salilab.org/publications/ref/ http://www.salilab.org/publications/papers/molmed-95/.

Sali, A; Potterton, L; Yuan, F; van Vlijmen, H; Karplus, M. 1995 Nov. Evaluation of comparative protein modeling by MODELLER. *PROTEINS: Structure, Function, and Genetics* 23(3):318-326.

Sali, A. 2001. MODELLER, 6a. Rockefeller University. http://www.salilab.org/.

Sanchez, R; Sali, A. 1997a Apr. Advances in comparative protein-structure modeling. *Current Opinion in Structural Biology* 7(2):206-214. http://www.salilab.org/publications/papers/cosb-96/html/ http://www.salilab.org/publications/ref/.

Sanchez, R; Sali, A. 1997b. Evaluation of comparative protein structure modeling by MODELLER-3. *PROTEINS: Structure, Function, and Genetics* Suppl 1:50-58.

Sanderson, M J; Baldwin, B G; Bharathan, G; Campbell, C S; von Dohlen, C; Ferguson, D; Porter, J M; Wojciechowski, M F; Donoghue, M J. 1993 Dec. The growth of phylogenetic information and the need for a phylogenetic data base. *Systematic Biology* 42(4):562-568. http://www.treebase.org.

Sanderson, M J. 1995 Sep. Objections to bootstrapping phylogenies: A critique. *Systematic Biology* 44(3):299-320.

Sandler, S J; Satin, L H; Samra, H S; Clark, A J. 1996 Jun 1. RecA-like genes from three archaean species with putative protein products similar to Rad51 and Dmc1 proteins of the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Research* 24(11):2125-2132. http://nar.oupjournals.org/cgi/content/abstract/24/11/2125.

Saqi, M A S; Russell, R B; Sternberg, M J E. 1998 Aug. Misleading local sequence alignments: Implications for comparative protein modeling. *Protein Engineering* 11(8):627-630. http://peds.oupjournals.org/cgi/content/abstract/11/8/627.

Saraf, M C; Moore, G L; Maranas, C D. 2003 Jun 1. Using multiple sequence correlation analysis to characterize functionally important protein regions. *Protein Engineering* 16(6):397-406.

Sauder, J M; Arthur, J W; Dunbrack, R L, Jr. 2000 Jul 1. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *PROTEINS: Structure, Function, and Genetics* 40(1):6-22.

Sawada, I; Schmid, C W. 1986 Dec 20. Primate evolution of the alpha-globin gene cluster and its *Alu*-like repeats. *Journal of Molecular Biology* 192(4):693-709.

Sayers, D L. 1934. *The Nine Tailors*. Gollancz. http://en.wikipedia.org/wiki/Change_ringing.

Schaffer, A A; Aravind, L; Madden, T L; Shavirin, S; Spouge, J L; Wolf, Y I; Koonin, E V; Altschul, S F. 2001 Jul 15. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research* 29(14):2994-3005. http://nar.oupjournals.org/cgi/content/full/29/14/2994.

Schmidt, H A; Strimmer, K S; Vingron, M; von Haeseler, A. 2002 Mar. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18(3):502-504. http://bioinformatics.oupjournals.org/cgi/content/abstract/18/3/502 http://www.tree-puzzle.de http://www.stat.uni-muenchen.de/~strimmer/cv.html.

Schwartz, R; King, J. 2006 Jan. Frequencies of hydrophobic and hydrophilic runs and alternations in proteins of known structure. *Protein Science* 15(1):102-112. http://www.proteinscience.org/cgi/content/abstract/15/1/102.

Schwede, T; Kopp, J; Guex, N; Peitsch, M C. 2003 Jul 1. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Research* 31(13):3381-3385. http://nar.oxfordjournals.org/cgi/content/full/31/13/3381.

Shafer, K S; Hanekamp, T; White, K H; Thorsness, P E. 1999 Oct. Mechanisms of mitochondrial escape to the nucleus in the yeast *Saccharomyces cerevisiae*. *Current Genetics* 36(4):183-194.

Shallom, S; Zhang, K; Jiang, L; Rathod, P K. 1999 Dec 31. Essential protein-protein interactions between *Plasmodium falciparum* thymidylate synthase and dihydrofolate reductase domains. *Journal of Biological Chemistry* 274(53):37781-37786.

Shi, Y; Yokoyama, S. 2003 Jul 8. Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates. *Proceedings of the National Academy of Sciences USA* 100(14):8308-8313. http://www.pnas.org/cgi/content/full/100/14/8308.

Shortle, D. 2002 Jan. Composites of local structure propensities: Evidence for local encoding of long-range structure. *Protein Science* 11(1):18-26. http://www.proteinscience.org/cgi/content/full/11/1/18.

Simmons, M P; Randle, C P; Freudenstein, J V; Wenzel, J W. 2002 Jan 1. Limitations of Relative Apparent Synapomorphy Analysis (RASA) for measuring phylogenetic signal. *Molecular Biology and Evolution* 19(1):14-23. http://mbe.oupjournals.org/cgi/content/abstract/19/1/14.

Simon, K V; Simon-Lukasik, K V. 1998. *Development of a putative structure of the catalytic domain of Microbispora bispora Endoglucanase A by homology modeling*. Mabel Smith Douglass Honors Program. Rutgers University, Douglass College (New Brunswick, NJ): Department of Biochemistry and Microbiology.

Singer, M S; Vriend, G; Bywater, R P. 2002 Sep 1. Prediction of protein residue contacts with a PDB-derived likelihood matrix. *Protein Engineering* 15(9):721-725. http://peds.oxfordjournals.org/cgi/content/full/15/9/721.

Smiley, J A; Paneth, P; O'Leary, M H; Bell, J B; Jones, M E. 1991 Jun 25. Investigation of the enzymatic mechanism of yeast orotidine-5'-monophosphate decarboxylase using 13C kinetic isotope effects. *Biochemistry* 30(25):6216-6223.

Smit, A F A. 1999 Dec 1. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics & Development* 9(6):657-663.

Smith, A W; Kahn, P C. 2005 Apr 2. Phylogenetics and homology modeling. *New Jersey Academy of Science and Affiliated Societies Annual Meeting* 50. New Jersey Institute of Technology, Newark, New Jersey.

Sneath, P H A; Sokal, R R. 1973. *Numerical taxonomy: The principles and practice of numerical classification*. W H Freeman and Company (San Francisco). ISBN 0-7167-0697-0.

Sommer, S S. 1992 Jul. Assessing the underlying pattern of human germline mutations: Lessons from the factor IX gene. *FASEB Journal* 6(10):2767-2774. http://www.fasebj.org/cgi/content/abstract/6/10/2767.

de Souza, P C; Bonilla-Rodriguez, G O. 2007 Jun. Fish hemoglobins. *Brazilian Journal of Medical and Biological Research* 40(6):769-778. http://dx.doi.org/10.1590/S0100-879X2007000600004.

Spencer, P. 1999 Apr 19. Applications of the Geometric Mean. University of Toronto. http://www.math.toronto.edu/mathnet/questionCorner/geomean.html http://en.wikipedia.org/wiki/Geometric_mean.

van der Spoel, D. 2002a Mar 16. [gmx-users] Re: xtal-water. http://www.gromacs.org/pipermail/gmx-users/2002-March/001044.html.

van der Spoel, D. 2002b Mar 15. [gmx-users] xtal-water. http://www.gromacs.org/pipermail/gmx-users/2002-March/001043.html.

van der Spoel, D; Lindahl, E; Hess, B; Groenhof, G; Mark, A E; Berendsen, H J C. 2005 Dec. GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* 26(16):1701-1718. http://folding.bmc.uu.se/ http://www.gromacs.org.

Spolsky, C; Uzzell, T. 1984 Sep 15. Natural interspecies transfer of mitochondrial DNA in amphibians. *Proceedings of the National Academy of Sciences USA* 81(18):5802-5805. http://www.pnas.org/cgi/content/abstract/81/18/5802.

Sproer, C; Mendrock, U; Swiderski, J; Lang, E; Stackebrandt, E. 1999 Oct 1. The phylogenetic position of *Serratia*, *Buttiauxella* and some other genera of the family Enterobacteriaceae. *International Journal of Systematic Bacteriology* 49(4):1433-1438. http://ijs.sgmjournals.org/cgi/content/abstract/49/4/1433.

Stallman, R; McGrath, R; Smith, P D. 1998. make, 3.77. Free Software Foundation. http://www.gnu.org/software/make/.

Stechmann, A; Cavalier-Smith, T. 2002 Jul 5. Rooting the eukaryote tree by using a derived gene fusion. *Science* 297(5578):89-91. http://www.sciencemag.org/cgi/content/full/297/5578/89.

Stechmann, A; Cavalier-Smith, T. 2003 Sep 2. The root of the eukaryote tree pinpointed. *Current Biology* 13(17):R665-R666.

Steel, M; Penny, D. 2000 Jun. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and Evolution* 17(6):839-850. http://mbe.oupjournals.org/cgi/content/full/17/6/839.

Sternberg, M J E; Bates, P A; Kelley, L A; MacCallum, R M. 1999 Jun. Progress in protein structure prediction: Assessment of CASP3. *Current Opinion in Structural Biology* 9(3):368-373.

Strimmer, K S; von Haeseler, A. 1996 Sep. Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution* 13(7):964-969. http://mbe.oupjournals.org/cgi/content/abstract/13/7/964 http://www.stat.uni-muenchen.de/~strimmer/cv.html.

Strimmer, K S. 1997. *Maximum likelihood methods in molecular phylogenetics*. Dissertation. University of Munich (Munich): Department of Biology. http://www.stat.uni-muenchen.de/~strimmer/cv.html.

Strimmer, K S; Goldman, N; von Haeseler, A. 1997 Feb. Bayesian probabilities and quartet puzzling. *Molecular Biology and Evolution* 14(2):210-211. http://mbe.oupjournals.org/cgi/content/abstract/14/2/210 http://www.stat.uni-muenchen.de/~strimmer/cv.html.

Strimmer, K S; von Haeseler, A. 1999. PUZZLE, 4.0.2. http://www.tree-puzzle.de ftp://ftp.ebi.ac.uk/pub/software/unix/puzzle/.

Stringer, J R. 1996 Oct. *Pneumocystis carinii*: What is it, exactly? *Clinical Microbiology Reviews* 9(4):489-498. http://cmr.asm.org/cgi/content/abstract/9/4/489.

Strych, U; Wohlfarth, S; Winkler, U K. 1994 Dec. Orotidine-5'-monophosphate decarboxylase from *Pseudomonas aeruginosa* PAO1: Cloning, overexpression, and enzyme characterization. *Current Microbiology* 29(6):353-359.

Suber, P. 2007 Jun 19. Open access overview: Focusing on open access to peer-reviewed research articles and their preprints. http://www.earlham.edu/~peters/fos/overview.htm.

Suchi, M. 1988 Mar. Molecular genetic studies on hereditary orotic aciduria I: Purification of human orotidine 5'-monophosphate decarboxylase and cloning of its DNA. *Nagoya Medical Journal* 32(3-4):207-220.

Sugita, T; Nakase, T. 1999a Feb. Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus *Candida*. *Systematic and Applied Microbiology* 22(1):79-86.

Sugita, T; Nakase, T. 1999b. Nonuniversal usage of the leucine CUG codon in yeasts: Investigation of basidiomycetous yeast. *The Journal of General and Applied Microbiology* 45(4):193-197. http://dx.doi.org/10.2323/jgam.45.193 http://www.jstage.jst.go.jp/article/jgam/45/4/45_193/_article.

Summa, C M; Levitt, M. 2007 Feb 27. Near-native structure refinement using *in vacuo* energy minimization. *Proceedings of the National Academy of Sciences USA* 104(9):3177-3182. http://www.pnas.org/cgi/content/full/104/9/3177.

Sung, P; Krejci, L; Van Komen, S; Sehorn, M G. 2003 Oct 31. Rad51 recombinase and recombination mediators. *Journal of Biological Chemistry* 278(44):42729-42732. http://www.jbc.org/cgi/content/full/278/44/42729.

Sunyaev, S R; Kuznetsov, E N; Rodchenkov, I V; Tumanyan, V G. 1997 Jun. Protein sequence-structure compatibility criteria in terms of statistical hypothesis testing. *Protein Engineering* 10(6):635-646. http://peds.oupjournals.org/cgi/content/abstract/10/6/635.

Sunyaev, S R; Eisenhaber, F; Argos, P; Kuznetsov, E N; Tumanyan, V G. 1998 May 15. Are knowledge-based potentials derived from protein structure sets discriminative with respect to amino acid types? *PROTEINS: Structure, Function, and Genetics* 31(3):225-246.

Swofford, D L. 1991. When are phylogeny estimations from molecular and morphological data incongruent? pp 294-333 in *Phylogenetic Analysis of DNA Sequences* ed: Miyamoto, M M; Cracraft, J. Oxford University Press (New York).

Taira, K; Benkovic, S J. 1988 Jan. Evaluation of the importance of hydrophobic interactions in drug binding to dihydrofolate reductase. *Journal of Medicinal Chemistry* 31(1):129-137.

Takahata, N; Slatkin, M. 1984 Mar 15. Mitochondrial gene flow. *Proceedings of the National Academy of Sciences USA* 81(6):1764-1767. http://www.pnas.org/cgi/content/abstract/81/6/1764.

Tarrio, R; Rodriguez-Trelles, F; Ayala, F J. 2000 Sep. Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: The *Drosophila saltans* and *willistoni* groups, a case study. *Molecular Phylogenetics and Evolution* 16(3):344-349.

Taylor, W R. 1994 Jun 30. Protein-structure modeling from remote sequence similarity. *Journal of Biotechnology* 35(2-3):281-291.

Taylor, W R; Flores, T P; Orengo, C A. 1994 Oct. Multiple protein structure alignment. *Protein Science* 3(10):1858-1870. http://www.proteinscience.org/cgi/content/abstract/3/10/1858.

Telford, M J; Herniou, E A; Russell, R B; Littlewood, D T J. 2000 Oct 10. Changes in mitochondrial genetic codes as phylogenetic characters: Two examples from the flatworms. *Proceedings of the National Academy of Sciences USA* 97(21):11359-11364. http://www.pnas.org/cgi/content/abstract/97/21/11359.

Telford, M J; Wise, M J; Gowri-Shankar, V. 2005 Apr. Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: Examples from the Bilateria. *Molecular Biology and Evolution* 22(4):1129-1136. http://mbe.oxfordjournals.org/cgi/content/full/22/4/1129.

Thompson, F L; Gevers, D; Thompson, C C; Dawyndt, P; Naser, S; Hoste, B; Munn, C B; Swings, J. 2005 Sep 1. Phylogeny and molecular identification of vibrios on the basis of multilocus sequence analysis. *Applied and Environmental Microbiology* 71(9):5107-5115.

Thompson, J D; Higgins, D G; Gibson, T J. 1994 Nov 11. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22(22):4673-4680.

Thompson, M J; Goldstein, R A. 1996 May. Constructing amino acid residue substitution classes maximally indicative of local protein structure. *PROTEINS: Structure, Function, and Genetics* 25(1):28-37.

Thompson, M J; Goldstein, R A. 1997 Sep. Predicting protein secondary structure with probabilistic schemata of evolutionarily derived information. *Protein Science* 6(9):1963-1975. http://www.proteinscience.org/cgi/content/abstract/6/9/1963.

Thorne, J L; Goldman, N; Jones, D T. 1996 May. Combining protein evolution and secondary structure. *Molecular Biology and Evolution* 13(5):666-673. http://mbe.oupjournals.org/cgi/content/abstract/13/5/666.

Thornton, J W; DeSalle, R. 2000. Gene family evolution and homology: Genomics meets phylogenetics. pp 41-73 ed: Lander, E S; Page, D; Lifton, R. Annual Review of Genomics and Human Genetics. Vol: 1. Annual Reviews (Palo Alto, CA).

Tillier, E R M; Collins, R A. 1995 Jan. Neighbor joining and maximum likelihood with RNA sequences: Addressing the interdependence of sites. *Molecular Biology and Evolution* 12(1):7-15. http://mbe.oupjournals.org/cgi/content/abstract/12/1/7.

Topham, C M; McLeod, A; Eisenmenger, F; Overington, J P; Johnson, M S; Blundell, T L. 1993 Jan 5. Fragment ranking in modeling of protein structure: Conformationally-constrained environmental amino acid substitution tables. *Journal of Molecular Biology* 229(1):194-220.

Traut, T W; Temple, B R. 2000 Sep 15. The chemistry of the reaction determines the invariant amino acids during the evolution and divergeance of orotidine 5'-monophosphate decarboxylase. *Journal of Biological Chemistry* 275(37):28675-28681. http://www.jbc.org/cgi/content/full/275/37/28675.

Tsai, J; Taylor, R; Chothia, C; Gerstein, M. 1999 Jul 2. The packing density in proteins: Standard radii and volumes. *Journal of Molecular Biology* 290(1):253-266. http://bioinfo.mbb.yale.edu/papers/.

Tuckwell, D S; Humphries, M J; Brass, A. 1995 Dec. Protein secondary structure prediction by the analysis of variation and conservation in multiple alignments. *Computer Applications in the Biosciences* 11(6):627-632.

Tuffery, P; Etchebest, C; Hazout, S. 1997 Apr. Prediction of protein side chain confirmations: A study on the influence of backbone accuracy on conformation stability in the rotamer space. *Protein Engineering* 10(4):361-372. http://peds.oupjournals.org/cgi/content/abstract/10/4/361.

Tuffley, C; Steel, M. 1998 Jan 1. Modeling the covarion hypothesis of nucleotide substitution. *Mathematics in the Biosciences* 147(1):63-91.

Turnbough, C L, Jr; Kerr, K H; Funderburg, W R; Donahue, J P; Powell, F E. 1987 Jul 25. Nucleotide sequence and characterization of the pyrF operon of *Escherichia coli K12*. *Journal of Biological Chemistry* 262(21):10239-10245. http://www.jbc.org/cgi/content/abstract/262/21/10239.

UniProt. 2005 Feb 1. SWISS-PROT, 46. Swiss Institute of Bioinformatics. http://www.expasy.org http://www.uniprot.org.

Van de Peer, Y; Frickey, T; Taylor, J S; Meyer, A. 2002 Aug 7. Dealing with saturation at the amino acid level: A case study based on anciently duplicated zebrafish genes. *Gene* 295(2 SI):205-211.

Vawter, L; Brown, W M. 1993 Jun. Rates and patterns of base change in the small subunit ribosomal RNA gene. *Genetics* 134(2):597-608. http://www.genetics.org/cgi/content/abstract/134/2/597.

Visiers, I; Braunheim, B B; Weinstein, H. 2000 Sep. Prokink: A protocol for numerical evaluation of helix distortions by proline. *Protein Engineering* 13(9):603-606. http://peds.oupjournals.org/cgi/content/abstract/13/9/603.

Vogt, G; Etzold, T; Argos, P. 1995 Jun 16. An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. *Journal of Molecular Biology* 249(4):816-831.

Voigt, C A; Gordon, D B; Mayo, S L. 2000 Jun 8. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *Journal of Molecular Biology* 299(3):789-803.

Voigt, C A; Mayo, S L; Arnold, F H; Wang, Z. 2001 Mar 27. Computational method to reduce the search space for directed protein evolution. *Proceedings of the National Academy of Sciences USA* 98(7):3778-3783. http://www.pnas.org/cgi/content/full/98/7/3778.

Vos, R. 2006. Bio:Phylo, 1.31. CPAN: Comprehensive Perl Archive Network. http://search.cpan.org/dist/Bio-Phylo/.

Wagner, P J. 2000. Phylogenetic analyses and the fossil record: Tests and inferences, hypotheses and models. *Paleobiology* 26(4 Suppl S):341-371.

Wako, H; Blundell, T L. 1994a May 19. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins I: Solvent accessibility classes. *Journal of Molecular Biology* 238(5):682-692.

Wako, H; Blundell, T L. 1994b May 19. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins II: Secondary structures. *Journal of Molecular Biology* 238(5):693-708.

Walden, K K O; Robertson, H M. 1997 Oct. Ancient DNA from amber fossil bees? *Molecular Biology and Evolution* 14(10):1075-1077. http://mbe.oupjournals.org/cgi/content/abstract/14/10/1075.

Wall, L; Christiansen, T; Orwant, J. 2000. *Programming Perl*. Edition: 3rd. O'Reilly (Sebastopol, CA). http://www.perl.org http://www.oreilly.com.

Wallner, B; Elofsson, A. 2003 May. Can correct protein models be identified? *Protein Science* 12(5):1073-1086. http://www.proteinscience.org/cgi/content/full/12/5/1073.

Wallqvist, A; Fukunishi, Y; Murphy, L R; Fadel, A; Levy, R M. 2000 Nov. Iterative sequence/secondary structure search for protein homologs: Comparison with amino acid sequence alignments and application to fold recognition in genome databases. *Bioinformatics* 16(11):988-1002. http://bioinformatics.oupjournals.org/cgi/content/abstract/16/11/988.

Wang, S-J; Chang, H-M; Lin, Y-S; Huang, C-H; Chen, C W. 1999 Sep 1. *Streptomyces* genomes: Circular genetic maps from the linear chromosomes. *Microbiology* 145(9):2209-2220.

Wang, Y; Anderson, J B; Chen, J; Geer, L Y; He, S; Hurwitz, D I; Liebert, C A; Madej, T; Marchler, G H; Marchler-Bauer, A; Panchenko, A R; Shoemaker, B A; Song, J S; Thiessen, P A; Yamashita, R A; Bryant, S H. 2002 Jan. MMDB: Entrez's 3D-structure database. *Nucleic Acids Research* 30(1):249-252. http://nar.oxfordjournals.org/cgi/content/full/30/1/249.

Wang, Y H; Bruenn, J A; Queener, S F; Cody, V. 2001 Sep. Isolation of rat dihydrofolate reductase gene and characterization of recombinant enzyme. *Antimicrobial Agents and Chemotherapy* 45(9):2517-2523.

Wanntorp, H-E. 1983. Reticulated cladograms and the identification of hybrid taxa. pp 81-88 in *Proceedings of the Second Meeting of the Willi Hennig Society* ed: Platnick, N I; Funk, V A. Advances in Cladistics. Vol: 2. Columbia University Press (New York).

Warshel, A; Strajbl, M; Villa, J; Florian, J. 2000 Dec 5. Remarkable rate enhancement of orotidine 5'-monophosphate decarboxylase is due to transition-state stabilization rather than to ground-state destabilization. *Biochemistry* 39(48):14728-14738.

Wayne, R K; Leonard, J A; Cooper, A. 1999. Full of sound and fury: The recent history of ancient DNA. pp 457-477 ed: Fautin, D G; Futuyma, D J; James, F C. Annual Review of Ecology and Systematics. Vol: 30. Annual Reviews (Palo Alto, CA).

Weiszfeld, E. 1937. Sur le point pour lequel la somme des distances de n points donnes est minimum. *Tohoku Mathematical Journal* 43:355-386. http://en.wikipedia.org/wiki/Geometric_median.

Wernegreen, J J; Moran, N A. 1999 Jan. Evidence for genetic drift in endosymbionts (*Buchnera*): Analyses of protein-coding genes. *Molecular Biology and Evolution* 16(1):83-97. http://mbe.oxfordjournals.org/cgi/content/abstract/16/1/83.

Westhead, D R; Collura, V P; Eldridge, M D; Firth, M A; Li, J; Murray, C W. 1995 Dec. Protein fold recognition by threading: Comparison of algorithms and analysis of results. *Protein Engineering* 8(12):1197-1204.

Wheeler, D L; Chappey, C; Lash, A E; Leipe, D D; Madden, T L; Schuler, G D; Tatusova, T A; Rapp, B A. 2000 Jan 1. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 28(1):15-18. http://nar.oxfordjournals.org/cgi/content/full/28/1/10.

Whelan, S; Goldman, N. 2001 May 1. A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. *Molecular Biology and Evolution* 18(5):691-699. http://mbe.oupjournals.org/cgi/content/full/18/5/691.

Wikipedia. 2006 Oct 4. Threading (protein sequence). Wikipedia. http://en.wikipedia.org/w/index.php?title=Threading_(protein_sequence)&oldid=84367928.

Wikipedia. 2008 Jan 30. Boltzmann distribution. http://en.wikipedia.org/wiki/Boltzmann_distribution.

Willson, S J. 2001 Mar. An error-correcting map for quartets can improve the signals for phylogenetic trees. *Molecular Biology and Evolution* 18(3):344-351. http://mbe.oupjournals.org/cgi/content/full/18/3/344.

Wilmanns, M; Eisenberg, D S. 1995 Jul. Inverse protein folding by the residue pair preference profile method: Estimating the correctness of alignments of structurally compatible sequences. *Protein Engineering* 8(7):627-639.

Wilson, K P; Malcolm, B A; Matthews, B W. 1992 May 25. Structural and thermodynamic analysis of compensating mutations within the core of chicken egg white lysozyme. *Journal of Biological Chemistry* 267(15):10842-10849. http://www.jbc.org/cgi/content/abstract/267/15/10842.

Winn, P J; Battey, J N D; Schleinkofer, K; Banerjee, A; Wade, R C. 2004 Feb 1. Issues in high-throughput comparative modelling: A case study using the ubiquitin E2 conjugating enzymes. *PROTEINS: Structure, Function, and Genetics* 58(2):367-375.

Wistrand, M. 2005. HMMer code alterations are tested for searches, not alignments. Email communication. To: Smith, A W.

Wistrand, M; Sonnhammer, E L L. 2005 Apr 15. Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMer. *BioMedCentral Bioinformatics* 6:99. ftp://ftp.cgb.ki.se/pub/prog/SAM_HMMER http://www.biomedcentral.com/1471-2105/6/99.

Wohlfahrt, G; Hangoc, V; Schomburg, D. 2002 May 15. Positioning of anchor groups in protein loop prediction: The importance of solvent accessibility and secondary structure elements. *PROTEINS: Structure, Function, and Genetics* 47(3):370-378.

Word, J M. 1999. AtVol, 1.2. Duke University (Durham, North Carolina). http://kinemage.biochem.duke.edu/software/utilities.php.

Word, J M; Lovell, S C; LaBean, T H; Taylor, H C; Zalis, M E; Presley, B K; Richardson, J S; Richardson, D C. 1999a Jan 29. Visualizing and quantifying molecular goodness-of-fit: Small-probe contact dots with explicit hydrogen atoms. *Journal of Molecular Biology* 285(4):1711-1733. http://kinemage.biochem.duke.edu/software/probe.php.

Word, J M; Lovell, S C; Richardson, J S; Richardson, D C. 1999b Jan 29. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology* 285(4):1735-1747. http://kinemage.biochem.duke.edu/software/reduce.php.

Word, J M. 2000. dang, 1.2.5. Duke University (Durham, North Carolina). http://kinemage.biochem.duke.edu/software/dang.php.

Word, J M; Bateman, R C; Presley, B K; Lovell, S C; Richardson, D C. 2000 Nov. Exploring steric constraints on protein mutations using MAGE/PROBE. *Protein Science* 9(11):2251-2259. http://www.proteinscience.org/cgi/content/full/9/11/2251.

Word, J M; Richardson, D C. 2006. reduce, 3.03. Duke University (Durham, North Carolina). http://kinemage.biochem.duke.edu/software/reduce.php.

Wroe, R; Al-Chalabi, A. 2007. Alsod: The ALS online database. The Institute of Psychiatry. http://alsod.iop.kcl.ac.uk/Als/.

Wu, N; Christendat, D; Dharamsi, A; Pai, E F. 2000a Jul. Purification, crystallization, and preliminary X-ray study of orotidine 5'-monophosphate decarboxylase. *Acta Crystallographica Section D* 56(7):912-914.

Wu, N; Mo, Y; Gao, J; Pai, E F. 2000b Feb 29. Electrostatic stress in catalysis: Structure and mechanism of the enzyme orotidine monophosphate decarboxylase. *Proceedings of the National Academy of Sciences USA* 97(5):2017-2022. http://www.pnas.org/cgi/content/full/97/5/2017.

Wu, T D; Nevill-Manning, C G; Brutlag, D L. 1999 Summer. Minimal-risk scoring matrices for sequence analysis. *Journal of Computational Biology* 6(2):219-235. http://cmgm.stanford.edu/~brutlag/Publications.html.

Wyss, A R; Novacek, M J; McKenna, M C. 1987 Mar. Amino acid sequence versus morphological data and the interordinal relationships of mammals. *Molecular Biology and Evolution* 4(2):99-116. http://mbe.oupjournals.org/cgi/content/abstract/4/2/99.

Xia, X; Xie, Z; Kjer, K M. 2003 Jun. 18S ribosomal RNA and tetrapod phylogeny. *Systematic Biology* 52(3):283-295.

Xu, S. 2000 Jun. Phylogenetic analysis under reticulate evolution. *Molecular Biology and Evolution* 17(6):897-907. http://mbe.oupjournals.org/cgi/content/full/17/6/897.

Yablonski, M J; Pasek, D A; Han, B-D; Jones, M E; Traut, T W. 1996 May 3. Intrinsic activity and stability of bifunctional human UMP synthase and its two separate catalytic domains, orotate phosphoribosyltransferase and orotidine-5'-phosphate decarboxylase. *Journal of Biological Chemistry* 271(18):10704-10708. http://www.jbc.org/cgi/content/full/271/18/10704.

Yang, A-S; Honig, B. 2000a Aug 18. An integrated approach to the analysis and modeling of protein sequences and structures I: Protein structural alignment and a quantitative measure for protein structural distance. *Journal of Molecular Biology* 301(3):665-678.

Yang, A-S; Honig, B. 2000b Aug 18. An integrated approach to the analysis and modeling of protein sequences and structures II: On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *Journal of Molecular Biology* 301(3):679-689.

Yang, A-S; Honig, B. 2000c Aug 18. An integrated approach to the analysis and modeling of protein sequences and structures III: A comparative study of sequence conservation in protein structural families using multiple structural alignments. *Journal of Molecular Biology* 301(3):691-711.

Yang, A-S. 2002 Dec 1. Structure-dependent sequence alignment for remotely related proteins. *Bioinformatics* 18(12):1658-1665. http://bioinformatics.oupjournals.org/cgi/content/abstract/18/12/1658.

Yang, W Z; Ko, T P; Corselli, L; Johnson, R C; Yuan, H S. 1998 Sep. Conversion of a beta-strand to an alpha-helix induced by a single-site mutation observed in the crystal structure of Fis mutant Pro(26)Ala. *Protein Science* 7(9):1875-1883.

Yang, Z. 1994 Sep. Statistical properties of the maximum-likelihood method of phylogenetic estimation and comparison with distance-matrix methods. *Systematic Biology* 43(3):329-342. http://abacus.gene.ucl.ac.uk/ziheng/cv.html.

Yang, Z; Kumar, S S C; Nei, M. 1995 Dec. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141(4):1641-1650. http://abacus.gene.ucl.ac.uk/ziheng/cv.html http://www.genetics.org/cgi/content/abstract/141/4/1641.

Yang, Z. 1996a Sep. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution* 11(9):367-372. http://abacus.gene.ucl.ac.uk/ziheng/cv.html.

Yang, Z. 1996b Feb. Phylogenetic analysis using parsimony and likelihood methods. *Journal of Molecular Evolution* 42(2):294-307. http://abacus.gene.ucl.ac.uk/ziheng/cv.html.

Yang, Z; Nielsen, R; Hasegawa, M. 1998 Dec. Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution* 15(12):1600-1611. http://mbe.oupjournals.org/cgi/content/abstract/15/12/1600 http://abacus.gene.ucl.ac.uk/ziheng/cv.html.

Yang, Z. 2000a Jan 22. Complexity of the simplest phylogenetic estimation problem. *Proceedings of the Royal Society of London B* 267(1439):109-116. http://abacus.gene.ucl.ac.uk/ziheng/cv.html http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=10687814.

Yang, Z. 2000b Nov. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *Journal of Molecular Evolution* 51(5):423-432. http://abacus.gene.ucl.ac.uk/ziheng/cv.html.

Yang, Z; Nielsen, R; Goldman, N; Pedersen, A-M K. 2000 May. Codon-substitution models for heterogenous selection pressure at amino acid sites. *Genetics* 155(1):431-449. http://abacus.gene.ucl.ac.uk/ziheng/cv.html http://www.genetics.org/cgi/content/full/155/1/431.

Yaoi, T; Laksanalamai, P; Jiemjit, A; Kagawa, H K; Alton, T; Trent, J D. 2000 Sep 7. Cloning and characterization of ftsZ and pyrF from the archaeon *Thermoplasma acidophilum*. *Biochemical & Biophysical Research Communications* 275(3):936-945.

Yona, G; Levitt, M. 2002 Feb 1. Within the twilight zone: A sensitive profile-profile comparison tool based on information theory. *Journal of Molecular Biology* 315(5):1257-1275.

Yumoto, I; Yamazaki, K; Sawabe, T; Nakano, K; Kawasaki, K; Ezura, Y; Shinano, H. 1998 Apr 1. *Bacillus horti* sp. nov., a new gram-negative alkaliphilic bacillus. *International Journal of Systematic Bacteriology* 48(2):565-571. http://ijs.sgmjournals.org/cgi/content/abstract/48/2/565.

Zajac, B. 1999. Math:Interpolate, 1.05. CPAN: Comprehensive Perl Archive Network. http://www.perl.com/CPAN/authors/id/B/BZ/BZAJAC/.

Zemla, A; Venclovas, C; Reinhardt, A; Fidelis, K; Hubbard, T J P. 1997. Numerical criteria for the evaluation of *ab initio* predictions of protein structure. *PROTEINS: Structure, Function, and Genetics* Suppl 1:140-150.

Zemla, A; Venclovas, C; Moult, J; Fidelis, K. 1999. Processing and analysis of CASP3 protein structure predictions. *PROTEINS: Structure, Function, and Genetics* Suppl 3:22-29.

Zhang, C; Kim, S-H. 2000 Mar 14. Environment-dependent residue contact energies for proteins. *Proceedings of the National Academy of Sciences USA* 97(6):2550-2555. http://www.pnas.org/cgi/content/full/97/6/2550.

Zhang, J; Nei, M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *Journal of Molecular Evolution* 44(Suppl 1):S139-S146.

Zhang, J; Rosenberg, H F. 2002 Apr 16. Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. *Proceedings of the National Academy of Sciences USA* 99(8):5486-5491. http://www.pnas.org/cgi/content/full/99/8/5486.

Zhang, J. 2003 Jul 8. Paleomolecular biology unravels the evolutionary mystery of vertebrate UV vision. *Proceedings of the National Academy of Sciences USA* 100(14):8045-8047. http://www.pnas.org/cgi/content/full/100/14/8045.

Zhang, K Y J; Eisenberg, D S. 1994 Apr. The three-dimensional profile method using residue preference as a continuous function of residue environment. *Protein Science* 3(4):687-695.

Zhang, N F. 2006 Jun. The uncertainty associated with the weighted mean of measurement data. *Metrologia* 43(3):195-204.

Zhang, Y P; Kolinski, A; Skolnick, J. 2003 Aug. TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophysical Journal* 85(2):1145-1164. http://www.biophysj.org/cgi/content/full/85/2/1145.

Zhang, Z; Gerstein, M. 2003 Sep 15. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Research* 31(18):5338-5348. http://nar.oxfordjournals.org/cgi/content/full/31/18/5338.

Zhao, H; Chen, M-H; Shen, Z-M; Kahn, P C; Lipke, P N. 2001 Jun. Environmentally induced reversible conformational switching in the yeast cell adhesion protein alpha-agglutinin. *Protein Science* 10(6):1113-1123. http://www.proteinscience.org/cgi/content/full/10/6/1113.

Zhou, H; Zhou, Y. 2005 Sep 15. SPEM: Improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics* 21(18):3615-3621. http://bioinformatics.oxfordjournals.org/cgi/content/full/21/18/3615.

Zhu, G; Keithly, J S; Philippe, H. 2000 Jul. What is the phylogenetic position of *Cryptosporidium*? *International Journal of Systematic and Evolutionary Microbiology* 50(4):1673-1681. http://ijs.sgmjournals.org/cgi/content/abstract/50/4/1673.

Zou, J M; Saven, J G. 2000 Feb 11. Statistical theory of combinatorial libraries of folding proteins: Energetic discrimination of a target structure. *Journal of Molecular Biology* 296(1):281-294.

Zou, J M; Saven, J G. 2003 Feb 22. Using self-consistent fields to bias Monte Carlo methods with applications to designing and sampling protein sequences. *Journal of Chemical Physics* 118(8):3843-3854.

Zrzavy, J. 2001. The interrelationships of metazoan parasites: A review of phylum- and higher-level hypotheses from recent morphological and molecular phylogenetic analyses. *Folia Parasitologica (Praha)* 48(2):81-103.

Curriculum Vita

Allen Watkins Smith

Education

Undergraduate

- Earlham College (Richmond, IN)
  - Dates attended: 9/1988-9/1992 then[565] 1/1994-6/1994
  - Major: Biology
  - Degree awarded (6/1994): Bachelor of Arts[566]
- University of South Alabama (Mobile, AL)
  - Dates attended[567]: 7/1990-9/1990, 7/1991-9/1991, 9/1993-11/1993, 9/1994-6/1995
  - Major/Degree: Not applicable (Nondegree)

Graduate

- Rutgers University and UMDNJ (Joint Program in Molecular Biosciences)
  - Dates attended: 9/1995-12/2007
  - Program: Microbiology and Molecular Genetics
  - Degree awarded (prospective; 1/2008): Doctor of Philosophy

Occupations

- Teaching Assistant
  - Dates: 9/1996-7/1997, 2/1998-7/2007
  - Department: Biochemistry and Microbiology, Cook College
  - Course: Experimental Biochemistry (2-semester course)
  - Professor: Dr. Theodore Chase, except for the 1999-2000 academic year, for which the professor was Dr. Theodorus van Es.
- Graduate Assistant
  - Dates: 9/1997-1/1998
  - Professor: Dr. Peter C. Kahn
  - Subject: Emulsions in food products

Positions

- System Administrator, Structural Biology Computational Laboratory
  - Dates: 8/1997-7/2000
  - Location: Lipman Hall Room 202, Cook Campus

---

[565] The gap is due to a medical leave of absence.

[566] A BA is the only degree granted by Earlham (other than via the affiliated Earlham School of Religion, a seminary); it is a liberal-arts college.

[567] Earlham does not have a summer session; attendance at the University of South Alabama was during the summer or the medical leave of absence.

- Fallback[568] System Administrator, Structural Biology Computational Laboratory
  - Dates: 8/2000-1/2008
  - Location: Lipman Hall Room 202, Cook Campus

Publications

Smith, D L; Smith, E A[569]. 1991 Mar 10-13. Decreased serum calcium and phosphate levels in older cystic fibrosis patients. *Annual Conference on Pulmonary Rehabilitation and Home Ventilation* 3. Poster Symposium. Denver, Colorado.

Smith, D L; Smith, E A; Sindel, L J. 1991. Differences in reactivity to commercial oak pollen extracts. *Annals of Allergy* 66(1):92.

Smith, E A; Smith, D L; Michaelski, J P. 1992. Cost-effective method of determining HLA-DQB1 genotypes. *Clinical Research* 40(4):817A.

Smith, E A; Seickel, J M; Smith, D L. 1992. Trends in skin test results in a southeastern community. *Annals of Allergy* 68(1):103.

Smith, E A; Smith, D L; Michaelski, J P. 1993. Allergic responsiveness and HLA type. *Journal of Allergy and Clinical Immunology* 91(1 part 2):339.

Smith, E A; Smith, D L; Michaelski, J P. 1993 Jan. Cost-effective method of determining HLA-DQB1 genotypes. *Southern Society for Clinical Research* Oral Presentation. New Orleans, Louisiana.

Smith, A W. 2001 May 29. What's the difference? *Science*. 292(5502)**:**E-Letter. http://www.sciencemag.org/cgi/eletters/292/5520/1303.

Smith, A W; Kahn, P C. 2005 Apr 2. Phylogenetics and homology modeling. *New Jersey Academy of Science and Affiliated Societies Annual Meeting* 50. New Jersey Institute of Technology, Newark, New Jersey.

---

[568] By “fallback” is meant, e.g., for emergencies, after normal work hours, and on weekends. This change was done to increase the time available for research, coursework, and teaching, particularly in consideration of the temporal load from the last.

[569] Please note a legal name change from Ed Allen Smith (Smith, E A) to Allen Watkins Smith (Smith, A W).