Inferred thermophily of the last universal ancestor based on estimated amino acid composition

Dawn J. Brooks and Eric A. Gaucher

Address: Foundation for Applied Molecular Evolution, Gainesville, Florida 32601

Telephone: (352) 271-7005

Fax: (352) 271-7076

Email address: dbrooks@ffame.org, egaucher@ffame.org

Key words: Last Universal Ancestor, optimal growth temperature, thermophile, amino acid composition, protein evolution, expectation maximization

Running head: Inferred thermophily of last universal ancestor

Abbreviations: LUA, last universal ancestor; OGT, optimal growth temperature; EM, expectation maximization; NJ, neighbor-joining

Abstract

The environmental temperature of the last universal ancestor (LUA) of all extant organisms is the subject of heated debate. Because the amino acid composition of proteins differs between mesophiles and thermophiles, the inferred amino acid composition of proteins in the LUA could be used to classify it as one or the other. We applied expectation maximization (EM) to estimate the amino acid composition of a set of thirty-one proteins in the LUA based on alignments of their modern day descendants, a phylogenetic tree relating those descendants and a model of evolution. Separate estimates of amino acid composition in LUA proteins were derived using modern day sequences of eight mesophilic species, eight thermophilic species or the sixteen species combined. We show that the relative mean Euclidean distance between the amino acid composition in one species and that of a set of mesophiles or thermophiles can be employed as a classifier with 100% accuracy. Applying this classifier to the estimated amino acid composition of the ancestral protein set in the LUA, we find it to be classified as a thermophile even when only the proteins of mesophilic species are used to derive the estimate. Based on the estimated amino acid composition of proteins in the LUA, we infer that it was a thermophile. We discuss our findings in the context of previous data pertaining to the OGT of the LUA, particularly the inferred G + C content of its rRNA. We conclude that the gathering evidence strongly supports a thermophilic LUA.

Introduction

The LUA represents a relatively accessible theoretical intermediary between extant cellular organisms and early, pre-cellular " life". Through analysis of modern day genomes it is possible to infer characteristics of the LUA (Lazcano and Forterre 1999) and these provide important clues to early evolution. One feature that has attracted significant interest is the temperature of the environment in which it lived (Woese 1987; Galtier, Tourasse, and Gouy 1999; DiGiulio 2001; Bocchetta et al. 2000; Brochier and Philippe 2002; Whitfield 2004;). The earliest evidence, from analysis of a phylogenetic reconstruction of the tree of life, prompted the proposal that the LUA was a thermophile, i.e., that it lived at temperatures > 55°C (Woese 1987); however, the reconstruction of the tree, and thus this evidence, remains controversial (Bocchetta et al. 2000; Brochier and Philippe 2002). Although it has been asserted that the inferred G + C content of ribosomal RNA in the LUA supports a mesophilic lifestyle (Galtier, Tourasse, and Gouy 1999), that claim does not hold up under scrutiny (Eric Gaucher, personal communication; see Discussion). Characteristics of ancestral proteins have also been brought to bear on this question. Experimentally reconstructed EF-Tu proteins of early ancestral bacteria were found to have temperature optima falling between 55 and 65°C, implying that those ancestors were thermophilic in nature (and by the most parsimonious extension, that the LUA was also) (Gaucher et al. 2003). In addition, the amino acid composition of the inferred sequences of signal recognition particle and a tRNA synthetase in the LUA was reported to be more similar to that of extant thermophiles than mesophiles (DiGiulio 2001). Because of the contradictory

nature of the evidence regarding the OGT of the LUA, a fresh approach that provides alternative data should help advance the debate.

There have been several studies reporting a relationship between OGT and amino acid composition (Kreil and Ouzounis 2001; Tekaia, Yeramian, and Dujon 2002; Singer and Hickey 2003; Nakashima, Fukuchi, and Nishikawa 2003). Taking advantage of this relationship, inferred amino acid composition of proteins in the LUA could be used to infer whether it was a thermophile or a mesophile. Using an approach analogous to that used by Galtier et al. (1999) to infer the ancestral G + C content of RNA, but addressing proteins rather than RNA, we previously estimated the amino acid composition of sixty-five proteins in the LUA and found it to be more similar to that of extant thermophiles than mesophiles (Brooks, Fresco, and Singh 2004). In the current analysis, we examined whether our previous result is robust with respect to the OGT of the taxa used to infer the amino acid composition of proteins in the LUA. We found that even if only mesophilic species are used to derive the estimated ancestral amino acid composition, that composition is most similar to that of thermophiles, as measured by Euclidean distance. We show that the relative mean Euclidean distance between the amino acid composition in any one species and that of a set of mesophiles or thermophiles can be used unequivocally to classify it. Thus, the inferred amino acid composition in the LUA allows us to classify it as a thermophile.

Methods

Included taxa and proteins

We sought to include orthologous proteins from as broad a phylogenetic distribution of genomes as possible while representing thermophiles and mesophiles equally in the data set.

Because it was important to utilize orthologs rather than paralogs in the analysis, and because orthologs can be difficult to distinguish from paralogs, we relied upon an established database, the Clusters of Orthologs Groups (COG) database (Tatusov et al. 2001) to aid the selection of proteins. Orthologous groups from the complete genomes from 30 major phylogenetic groups were available as of January 2004.

To be relatively confident that a protein family had been present in the LUA, we required two basic criteria be met. First, we required that members of the family be present in the clear majority of taxa (> 25). Second, we sought to exclude from the analysis protein families whose presence in the majority of taxa might be due to horizontal transfer between the primary lineages rather than to vertical inheritance. To meet this latter criterion, we required bacterial, archaeal and eukaryotic family members to form separate phylogenetic clades. In addition, we selected families in which the presence of paralogs would not confound the construction of a phylogenetic tree from the concatenated protein sequences, i.e., any paralogs had to be the result of post-speciation duplications, clustering as neighbors on the protein family tree.

Three criteria were used to select genomes for inclusion in the analysis. First, we sought an equal number of thermophiles and mesophiles. Second, inclusion of a genome should not dramatically reduce the number of shared orthologs meeting the criteria listed above. Third, we sought to represent the broadest possible phylogenetic distribution of taxa.

Using the COG database, seven thermophiles-two bacteria, *Aquifex aeolicus*, *Thermotoga maritima*, and five archaea, *Aeropyrum pernix*, *Thermoplasma acidophilum*, *Methanococcus jannaschii*, *Pyrococcus horikoshii* and *Archaeoglobus fulgidus*-were available for inclusion in our analysis (ignoring closely related taxa such as *P. horikoshii*

and *P. abyssi*). A phylogenetic tree was built for the thirty-four concatenated orthologs of the mesophilic taxa, and the seven which represented the greatest total branch lengths, and thus could be assumed to represent the greatest phylogenetic diversity, were selected-one eukaryote, *Saccharomyces cerevisiae*, and six bacteria, *Synechocystis*, *Xylella fastidiosa*, *Helicobacter pylori*, *Treponema pallidum*, *Chlamydia pnemoniae* and *Bacillus subtilis*. One mesophile, *Halobacterium sp. NRC-1*, was excluded because its inclusion would have reduced the number of sequences meeting our criteria from thirty-four to twenty-six.

Two additional taxa not available in the COG database were included in the analysis to increase the phylogenetic diversity of the individual mesophilic and thermophilic species sets. *Methanosarcina acetivorans* was included in order to have representation of a mesophilic archaean. Inclusion of *Thermoanaerobacter tengcongensis* allowed for representation of a thermophilic bacterium known not to be located basally in the bacterial lineage. (It is a member of *Firmicutes*, and therefore clusters with *B. subtilis*.) To identify the orthologs from *T. tengcongensis* and *M. acetivorans* belonging to each COG protein family, profile HMMs were built using the alignment of proteins collected for the fourteen COG taxa. These were then used to search the database of predicted protein sequences for each of the two genomes. Best hits were individually examined to ascertain that their annotation was consistent with the COG protein family. We were unable to identify orthologs for three COG families in both the additional taxa, so that our final set of ortholog families was reduced to thirty-one (Table 1).

The OGT of each species is listed in Table 2.

Alignments and phylogenetic trees

The program T-Coffee (Notredame, Higgins, and Heringa 2000) with default parameter settings was used to build alignments. Columns containing gaps were removed. Concatenation of the ungapped alignments resulted in a single alignment of 4449 residues. A phylogenetic tree was inferred using the neighbor-joining algorithm (Saitou and Nei 1987) as implemented in the Phylip software package (19), using its default parameter settings (Fig. 1a). The topology of the neighbor-joining tree was found to be congruent with a consensus tree for 100 bootstrap replicates (see Fig. 2), although the bootstrap support was as low as 59 for certain clusters within the bacterial lineage. The Bayesian phylogenetic inference software Mr. Bayes 3 (20) was also used to build a phylogenetic tree (Fig. 1b). Markov chain Monte Carlo resampling of tree parameters was performed with four chains and allowed to run for 150000 generations. A mixed model of amino acid substitution was used. Both NJ and Bayesian trees were midpoint rooted; however, analyses using alternative rootings of the sixteen-taxon tree, either at the base of the eukaryotic/archaeal divergence or the base of the bacterial divergence, led to identical conclusions as those using the midpoint-rooted tree. Alignments for the sequences of mesophilic and thermophilic taxa were extracted from the larger alignment of sixteen taxa (i.e., preserving the columns thereof). Similarly, trees for the mesophilic and thermophilic taxa were extracted from the sixteen-taxon tree, using the branch lengths and topology of that tree.

The EM implementation

We used the rate matrix of Jones et al. (21) as the model of evolution. A discrete gamma distribution was used to allow for rate variation between columns of the alignment, the rate

categories being estimated using the software package PAML (22). The program implementing the EM method is available from the author upon request.

Jackknife test of Euclidean distance as a classifier

For each of the sixteen taxa in turn, that test species was removed from the set of reference thermophilic and mesophilic species. The Euclidean distance between the amino acid composition of the set of thirty-one proteins within the test species and each of the reference species was calculated. The mean Euclidean distance between the test and the reference thermophilic species and the test and the reference mesophilic species was determined.

Results and Discussion

Sixteen fully sequenced genomes, including eight thermophiles and eight mesophiles and representing a broad phylogenetic distribution, were included in the analysis. (We did not make the distinction between thermophiles and hyperthermophiles, but instead grouped together all those species with OGT > 55°C as thermophiles.) In selecting a set of protein families for analysis, the set was restricted to those in which orthology between members of a family within the sixteen taxa was unambiguous and there was an absence of evidence for horizontal transfer between the primary lineages (bacteria, archaea and eukaryotes). Thirty-one proteins, the majority of them ribosomal, met these criteria. For a list of these proteins see Table 1.

Sequences of the thirty-one families were aligned and used to build a phylogenetic tree for the sixteen taxa using either neighbor-joining (NJ) or Bayesian phylogenetic methods

(Fig. 1). Difficulty resolving the bacterial lineages within the tree is suggested by weak bootstrap support for the branching order of much of the bacterial lineage in the NJ tree (Fig. 2) and by the fact that the two taxa, *Bacillus subtilis* and *Thermoanaerobacter tengcongensis*, belonging to the phylum *Firmicutes*, are not correctly clustered in the Bayesian tree. Of particular potential relevance to our investigation, because basal lineages (of moderate or short branch length) will have relatively strong influence on the estimated ancestral sequence, is whether the thermophilic species *Aquifex aeolicus* and *Thermotoga maritima* truly represent basal branches in the bacterial lineage. (The mesophile *Saccharomyces cerevisiae* is basal in the eukaryotic/archael lineages in both trees.) Because these two thermophilic species are basal in the Bayesian tree but not in the NJ tree, we employed an additional phylogenetic tree in which the bacterial taxa are related to each other by a star phylogeny, with branch lengths to the last common ancestor of the bacteria equal to the average distance to the bacterial divergence in the NJ tree and the eukaryotic/archaeal topology and branch lengths taken from the NJ tree. If anything, the star phylogeny is biased toward mesophilic species in terms of reconstruction of the ancestral sequence, because all taxa have equal influence on the ancestral sequence and six of the nine bacterial species are mesophiles. Because no outgroup exists for our phylogenies, trees were midpoint rooted; however, moving the root all the way to the base of the bacterial divergence or to the divergence of the archaea and eukaryotes does not result in qualitatively different results to those reported here.

Five estimates of the amino acid composition in the LUA were derived. For three, we used sequences of all sixteen taxa with the three alternative phylogenetic trees described above (NJ, Bayesian and NJ plus star). For a forth, we used only the sequences of the eight

mesophilic taxa, and for the last, we used only sequences of the eight thermophilic taxa. For the estimates using solely mesophilic or thermophilic taxa, phylogenetic trees that had the same topology and branch lengths of the sixteen-taxon NJ tree were assumed.

As with any EM approach (Dempster, Laird, and Rubin 1977), our method consists of an iteration of expectation and maximization steps. In the expectation step, the posterior probabilities of all twenty amino acids at the root node of a phylogenetic tree are derived for each position of the alignment, assuming their prior probabilities in the ancestral sequence and a model of evolution (Brooks, Fresco, and Singh 2004). From these, expected counts of each amino acid in the ancestral sequence can be calculated. In the maximization step, the frequency of each amino acid in the ancestral sequence is estimated as the expected counts of that amino acid in the reconstructed sequence divided by the length of the sequence. These new estimates of ancestral amino acid frequencies are used as the prior probabilities in the next expectation step, and the procedure is iterated to convergence (Brooks, Fresco, and Singh 2004). For estimates of amino acid frequencies in the LUA using the different sequence sets and phylogenetic trees see Table 3.

To determine whether the amino acid composition inferred in the LUA is more similar to that of extant mesophiles or thermophiles, we calculated the mean Euclidean distance between the estimated amino acid composition in the LUA and the composition observed within the thermophilic and mesophilic sequence sets used in the analysis. The mean Euclidean distance between the estimated LUA and the thermophilic amino acid composition is significantly smaller than the mean distance between the LUA and the mesophilic amino acid composition (P-value < 0.05) for all data sets. Significantly, this is the case even for LUA estimates derived using only mesophilic sequences (Fig. 3).

Using jackknife resampling, we examined whether the relative size of the mean Euclidean distance between the amino acid composition of the set of thirty-one proteins in one species and that of the mesophilic species and the mean Euclidean distance between the amino acid composition of the protein set in the same species and that of the thermophilic species could be used successfully to classify that species as a mesophile or thermophile, i.e., whether a species may be classified according to which set its amino acid composition is more similar to, as measured by Euclidean distance. We found this proposed classifier to have an accuracy of 100% (Fig. 3). Accordingly, based on the inferred amino acid composition of a set of thirty-one proteins in the LUA, the LUA can be classified unequivocally as a thermophile, even when proteins of modern day mesophiles alone are used to derive the estimate.

Our method for estimating amino acid composition of ancestral proteins is closely analogous to that of Galtier et al. (1999), in which EM was used to infer G+C content of ancestral rRNA sequences from extant ones. Those investigators, however, concluded that the inferred composition of rRNA in the LUA is inconsistent with its having been a thermophile. Because their findings are in direct contradiction to ours, we feel it is worthwhile to briefly discuss their data and analysis. It is apparent from the data presented in ref. 2 that there is no, or at most a very weak, correlation between OGT and rRNA G+C content for species with OGT < 40°C. Consequently, G+C content cannot be a statistically sound means of predicting the OGT of a species. It is apparent from their data, in fact, that the inferred rRNA G+C content in the LUA is compatible with either a thermophilic or a mesophilic lifestyle.

Although we have focused here on using the estimated amino acid composition in the LUA to make inferences about the optimal growth temperature of the LUA, this composition may also provide additional clues to the early evolution of life, such as the establishment of the genetic code, a possibility that we have explored elsewhere (Brooks et al. 2002; Brooks, Fresco, and Singh 2004). Analysis of protein sequences of extant organisms may be a key, yet underutilized resource for constructing a more complete model of early life on this planet.

Conclusion

Our results provide strong support for the hypothesis that the LUA was a thermophile, i.e., that it lived at temperatures above 55°C. Using Euclidean distance as a measure, the estimated amino acid composition of proteins in the LUA is more similar to that of extant thermophiles than that of mesophiles, even when that estimate is derived using sequences solely from mesophilic species. We show using jackknife sampling that mean Euclidean distance of the protein amino acid composition of a species to the composition in a set of mesophilic or thermophilic species is 100% accurate as a classifier, choosing the set to which it is closest, and thus the LUA may be inferred to be a thermophile.

We note that the majority of data currently supports a proposed hot environment for the LUA. Approaches based on reconstruction of ancestral protein sequences consistently imply a thermophilic LUA (DiGiulio 2001; Gaucher et al. 2003; Brooks, Fresco, and Singh 2004; this study). Inferred G + C content of ribosomal RNA in the LUA does not exclude the possibility of a thermophilic LUA, as we briefly explained. Phylogeny-based inferences of ancestral environments have been divided in their conclusions (Bocchetta et al. 2000;

Brochier and Philippe 2002) and seem likely to remain so until agreement can be reached on the appropriate means for phylogeneic reconstruction of such anciently diverging lineages. Nonetheless, we are optimistic that with additional data and analyses, the remaining discrepancies between approaches will be resolved, leading ultimately to a consensus view.

Acknowledgments

We are grateful to Steve Benner for helpful discussions and to Phillip Prodger for providing comments of the manuscript. This work was supported by the National Science Foundation under a grant awarded in 2003 to DJB.

Literature cited

Bocchetta, M., S. Gribaldo, A. Sanangelantoni, and P. Cammarano. 2000. Phylogenetic depth of the bacterial genera *Aquifex* and *Thermotoga* inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. J. Mol. Evol. **50**: 366-380.

Brochier, C. and H. Philippe. 2002. Phylogeny: a non-hyperthermophilic ancestor for bacteria. Nature **417**: 244.

Brooks, D.J., J.R. Fresco, and M. Singh. 2004. A novel method for estimating ancestral amino acid composition and its application to proteins of the Last Universal Ancestor. Bioinformatics **20**:2251-2257.

Brooks, D.J., J.R. Fresco, A.M. Lesk, and M. Singh. 2002. Evolution of amino acid frequencies in proteins over deep time: Inferred order of introduction of amino acids into the genetic code. Mol. Biol. Evol. **19**, 1645-1655.

Dempster, A.P., N.M. Laird, D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Statist. Soc. B **39**: 1-38.

DiGiulio, M. 2001.The universal ancestor was a thermophile or a hyperthermophile. Gene **281**: 11-17.

Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 35c Distributed by the author. Department of Genetics, University of Washington, Seattle, 1993.

Galtier, N. and J.R. Lobry. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J. Mol. Evol. **44**:632-636.

Galtier, N., N. Tourasse, and M. Gouy. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science **283**: 220-221.

Gaucher, E.A., J.M. Thomson, M.F. Burgan, and S.A. Benner. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature **425**: 285-288.

Jones, D.T., W.R. Taylor, and J.M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**: 275-282.

Kreil, D.P. and C.A. Ouzounis. 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nuc. Acid Research **29**:1608-1615.

Lazcano, A. and P. Forterre. 1999. The molecular search for the last common ancestor. J. Mol. Evol. **49**: 411-412.

Nakashima, H., S. Fukuchi, and K. Nishikawa. 2003. Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. J. Biochem. **133**: 507-513.

Notredame, C., D.G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302**: 205-217.

Ronquist, F. and J.P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572-1574.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**: 406-425.

Singer, G.A.C. and D.A. Hickey. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene **317**: 39-47.

Tatusov, R.L., D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, and E.V. Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. **29**: 22-28.

Tekaia, F., E. Yeramian, and B. Dujon. 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene **297**: 51-60.

Whitfield, J. 2004. Born in a watery commune. Nature **427**: 674-676.

Woese, C.R. 1987. Bacterial evolution. Microbiol. Rev. **51**: 221-271.

Yang, Z. 1997. PAML: a program for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**: 555.

Table 1. List of COG database families included in analysis

COG ID	Protein name
COG0012	Predicted GTPase
COG0024	Methionine aminopeptidase
COG0048	Ribosomal protein S12
COG0051	Ribosomal protein S10
COG0052	Ribosomal protein S2
COG0080	Ribosomal protein L11
COG0081	Ribosomal protein L1
COG0087	Ribosomal protein L3
COG0088	Ribosomal protein L4
COG0090	Ribosomal protein L2
COG0091	Ribosomal protein L22
COG0092	Ribosomal protein S3
COG0093	Ribosomal protein L14
COG0097	Ribosomal protein L6
COG0098	Ribosomal protein S5
COG0100	Ribosomal protein S11
COG0102	Ribosomal protein L13
COG0103	Ribosomal protein S9
COG0180	Tryptophanyl-tRNA synthetase
COG0184	Ribosomal protein S15P/S13E
COG0185	Ribosomal protein S19
COG0186	Ribosomal protein S17
COG0197	Ribosomal protein L16/L10E
COG0199	Ribosomal protein S14
COG0200	Ribosomal protein L15
COG0201	Preprotein translocase subunit SecY
COG0244	Ribosomal protein L10
COG0250	Transcription antiterminator
COG0495	Leucyl-tRNA synthetase
COG0522	Ribosomal protein S4 and related proteins
COG0541	Signal recognition particle GTPase

Table 2. Species OGT in °C

Species	OGT
Saccharomyces cerevisiae	25
Synechocystis	25
Xylella fastidiosa	26
Bacillus subtilis	30
Chlamydia pnemoniae	37
Helicobacter pylori	37
Treponema pallidum	37
Methanosarcina acetivorans	40
Thermoplasma acidophilum	60
Thermoanaerobacter tengcongensis	75
Thermotoga maritima	80
Methanococcus jannaschii	82
Aquifex aeolicus	85
Archaeoglobus fulgidus	85
Aeropyrum pernix	90
Pyrococcus horikoshii	95

Table 3. Amino acid frequencies estimated in the LUA

The three data sets are indicated as All, all sixteen taxa; Thermo, the eight thermophilic taxa; and Meso, the eight mesophilic taxa. The alternative trees are indicated as NJ, the neighbor-joining tree; Bayes, the Bayesian tree; and NJ+star, the neighbor-joining tree with the bacteria assigned a star phylogeny. Amino acids are indicated by their three-letter code.

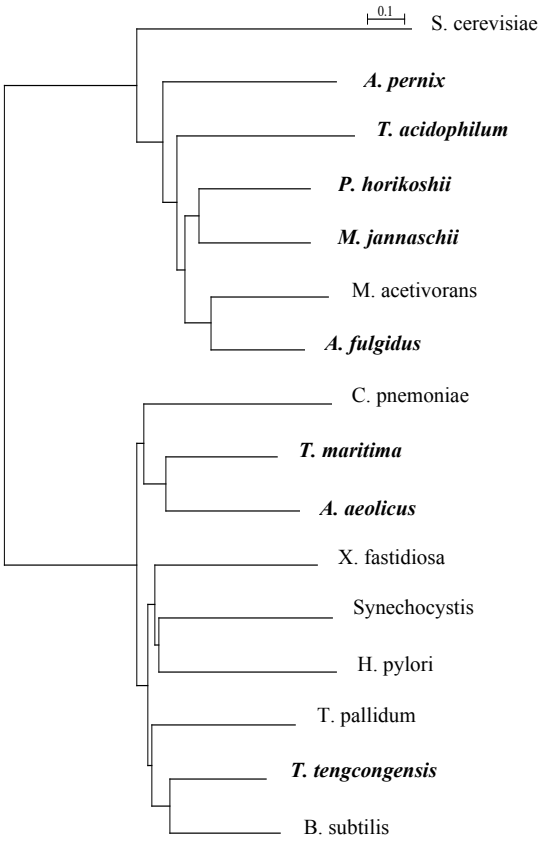
Set:	All	All	All	Thermo	Meso	mean	sd
Root:	NJ	Bayes	NJ+star	NJ	NJ		
Ala	0.0830	0.0831	0.0827	0.0840	0.0857	0.0837	0.0012
Arg	0.0874	0.0897	0.0868	0.0897	0.0844	0.0876	0.0022
Asn	0.0263	0.0239	0.0266	0.0239	0.0329	0.0267	0.0037
Asp	0.0354	0.0350	0.0356	0.0372	0.0369	0.0360	0.0010
Cys	0.0022	0.0027	0.0023	0.0027	0.0018	0.0023	0.0004
Gln	0.0153	0.0150	0.0155	0.0141	0.0184	0.0157	0.0016
Glu	0.0773	0.0770	0.0767	0.079	0.0637	0.0747	0.0062
Gly	0.0865	0.0858	0.0866	0.0846	0.0913	0.0870	0.0026
His	0.0217	0.0213	0.0217	0.0209	0.0222	0.0216	0.0005
Ile	0.1040	0.1031	0.1036	0.1046	0.0994	0.1029	0.0021
Leu	0.0706	0.0696	0.0708	0.0714	0.0722	0.0709	0.0010
Lys	0.1168	0.1182	0.1172	0.1118	0.1056	0.1139	0.0053
Met	0.0199	0.0185	0.0203	0.0194	0.0225	0.0201	0.0015
Phe	0.0261	0.0265	0.0263	0.0243	0.0277	0.0262	0.0012
Pro	0.0422	0.0438	0.0420	0.0445	0.0397	0.0424	0.0019
Ser	0.0232	0.0208	0.0236	0.0200	0.0325	0.0240	0.0050
Thr	0.0363	0.0365	0.0364	0.0369	0.0411	0.0374	0.0021
Trp	0.0016	0.0022	0.0017	0.0045	0.0011	0.0022	0.0013
Tyr	0.0176	0.0188	0.0175	0.0220	0.0140	0.0180	0.0029
Val	0.1063	0.1082	0.1059	0.1042	0.1067	0.1063	0.0014

Figure legends

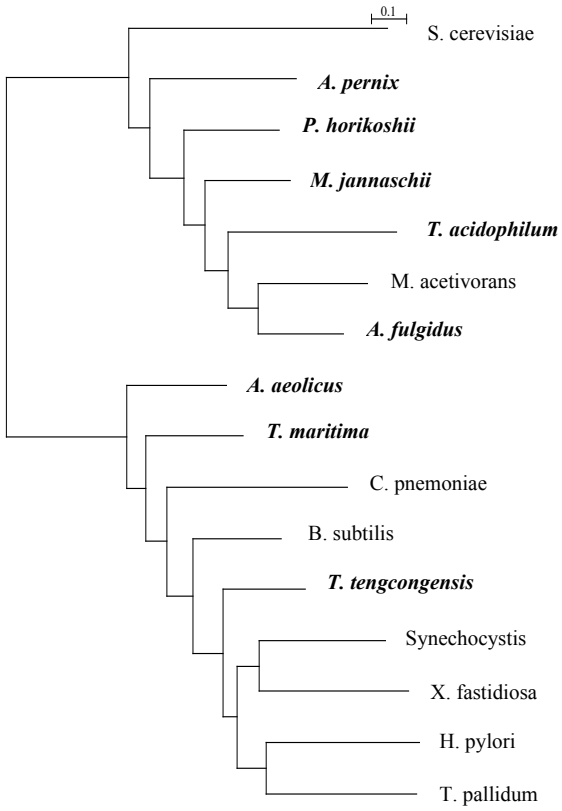
Figure 1. The phylogenetic trees used to derive EM estimates of ancestral amino acid composition using all sixteen taxa. Mesophiles are in normal text and thermophiles are in bold italics. Scale bars indicates 0.1 substitution per site. A. Neighbor-joining tree. Trees employed for estimates based solely on mesophilic or thermophilic taxa used the branch lengths and topology of the sixteen-taxon neighbor-joining tree. B. Bayesian tree.

Figure 2. Consensus 16-taxon NJ tree for 100 bootstrap replicates.

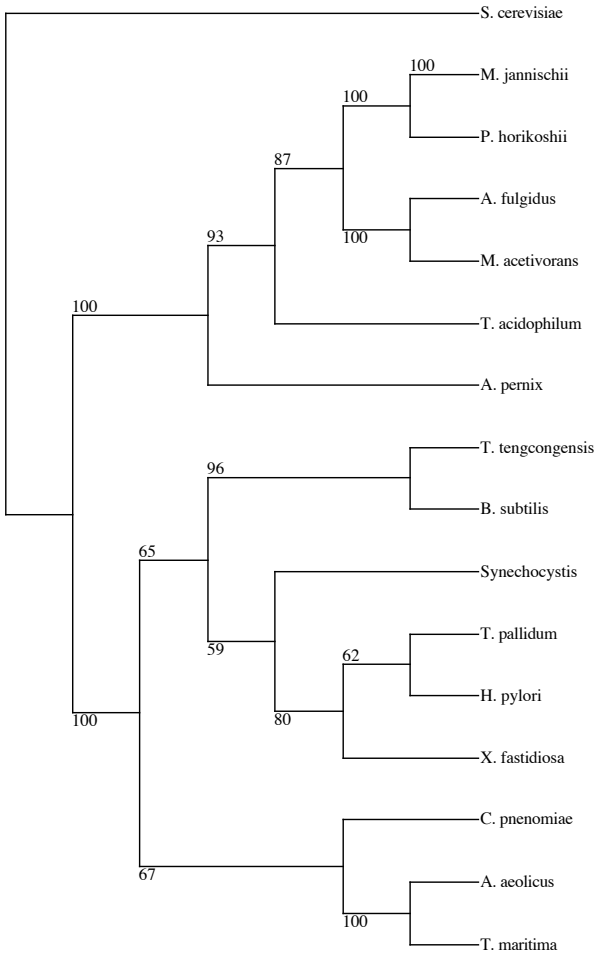
Figure 3. Jackknife data examining whether the relative average Euclidean distance between a species and either a set of mesophiles or a set of thermophiles may be used as a classifier. Each data point represents the mean Euclidean distance between the amino acid composition of the set of thirty-one proteins in a test species and the reference mesophilic (x axis) and thermophilic species (y axis). 'meso' indicates mesophilic test species, 'thermo' indicates thermophilic test species, and 'LUA-all', 'LUA-t', and 'LUA-m' indicate the estimated amino acid composition in the LUA using, respectively, all sixteen taxa, the eight thermophilic taxa, and the eight mesophilic taxa.



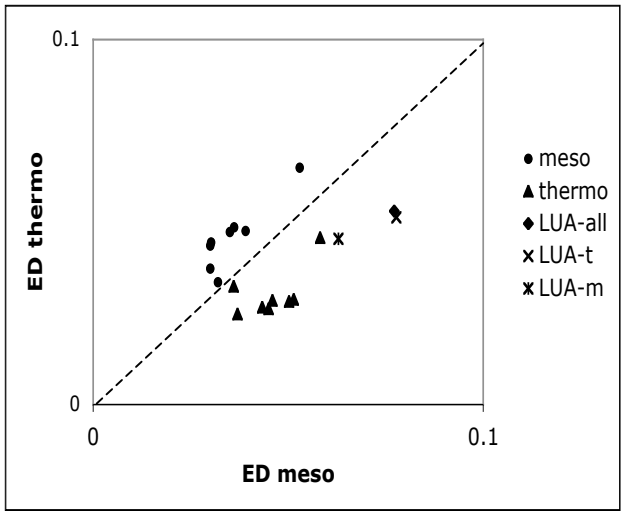
Brooks Figure 1a



Brooks Figure 1b



Brooks Figure 2



Brooks Figure 3