

**A novel method for estimating ancestral amino acid composition and its application
to proteins of the Last Universal Ancestor**

Brooks, D.J. Department of Genetics, Washington University School of Medicine, St
Louis, Missouri 63110 USA

Fresco, J.R. Department of Molecular Biology, Princeton University, Princeton, New
Jersey 08544 USA

Singh, M.* Department of Computer Science and the Lewis-Sigler Institute for
Integrative Genomics, Princeton University, Princeton, New Jersey 08544 USA

Running head: Estimation of ancestral amino acid composition

* To whom correspondence should be addressed.

Abstract

Motivation: Knowledge of how proteomic amino acid composition has changed over time is important for constructing realistic models of protein evolution and increasing our understanding of molecular evolutionary history. The proteomic amino acid composition of the Last Universal Ancestor of life (LUA) is of particular interest, since that might provide insight into the early evolution of proteins and the nature of the LUA itself.

Results: We introduce a method to estimate ancestral amino acid composition that is based on expectation-maximization (EM). On simulated data, the approach was found to be very effective in estimating ancestral amino acid composition, with accuracy improving as the number of residues in the dataset was increased. The method was then used to infer the amino acid composition of a set of proteins in the LUA. In general, as compared with the modern protein set, LUA proteins were found to be richer in amino acids that are believed to have been most abundant in the prebiotic environment and poorer in those believed to have been unavailable or scarce. Additionally, we found the inferred amino acid composition of this protein set in the LUA to be more similar to the observed composition of the same set in extant thermophilic species than in extant mesophilic species, supporting the idea that the LUA lived in a thermophilic environment.

Availability: The program is available upon request.

Contact: mona@cs.princeton.edu

Introduction

Amino acid composition is one of the most basic features of a proteome. Proteomic surveys have naturally included analysis of this feature (Gerstein, 1998), and as the number of proteomes available for analysis has grown, it has become evident that this feature displays considerable variation. (See, for example, the Proteome Analysis database (Pruess et al., 2003).) Perhaps the simplest, and best studied, explanation of this variation is the range in underlying genomic G+C content, a relationship that has been investigated over several decades (Sueoka, 1961; Knight et al, 2001). However, there are other determinants of proteomic amino acid composition, such as the optimal growth temperature of an organism (Kreil and Ouzounis, 2001; Saunders et al., 2003). Given the range of amino acid compositions observed and factors that influence amino acid composition, it is clear that this proteomic feature evolves.

We are concerned with the problem of inferring the amino acid composition of ancestral proteomes, or in practice, of protein subsets thereof. We are particularly interested in inferring the amino acid composition of a large protein set in the Last Universal Ancestor of all life (LUA). Firstly, this amino acid composition may indicate something about the nature of the LUA itself. For example, Galtier et al. (1999) used the inferred G+C content of the LUA ribosomal RNAs as evidence against its being thermophilic; and inferences about the amino acid composition of LUA proteins should also be pertinent to this particular question (Di Giulio, 2001). Furthermore, the amino acid composition of LUA proteins might provide clues to the early evolution of proteins and the origin of the genetic code (Brooks et al., 2002).

One approach for estimating the amino acid composition of an ancestral protein set is to infer the ancestral sequences of a contemporary protein set and then compute the composition of the inferred sequences (Di Giulio, 2001). However, existing methods for reconstructing ancestral protein sequences, maximum likelihood (ML) and maximum parsimony (MP), have limitations that compromise their ability to provide accurate estimates of ancestral sequence composition. Traditional ML methods require an assumption of the amino acid composition of the sequences being reconstructed, and it is usually assumed that the composition of the ancestors is the same as that of extant descendants (Yang et al., 1995). This requirement inherently limits the application of traditional ML methods to the problem of estimating ancestral amino acid composition. Although MP does not require any assumption regarding the composition of ancestral sequences, because it is less accurate at inferring ancestral sequences than ML (Yang et al., 1995), an ML-based method is preferable.

We have previously described a method for estimating ancestral amino acid composition that used MP to partially infer ancestral protein sequences in order to identify conserved residues in modern sequences. Following Bayes' Rule, the amino acid composition of such conserved residues, as well as the relative probability of conservation of each amino acid over the course of evolution, were used to estimate the amino acid composition of the full-length ancestral sequence set (Brooks et al., 2002).

Here we describe an alternate approach, based on ML sequence reconstruction (Yang et al., 1995), that uses expectation maximization (EM) (Dempster et al., 1977) to relax the requirement that the amino acid composition of the ancestral sequence be known a priori. Compared with the one we described previously, the new approach more fully

exploits models of amino acid substitution and makes estimates based on information present in both conserved and non-conserved residues. Although EM has been used to address other problems of sequence analysis and phylogenetic inference, the current application is entirely novel.

Galtier and Gouy (1998) previously described a ML approach for estimating the ancestral G+C content of a group of nucleotide sequences. Our approach is similar to theirs in some respects. However, Galtier and Gouy use the Newton-Raphson method to estimate several parameters of their model other than ancestral G+C frequency, including the equilibrium G+C content in each lineage. Because such an approach is unlikely to be applicable to protein sequences due to the substantially larger number of parameters required to be estimated, we make the simplifying assumption of constant equilibrium amino acid content in all descendant lineages (i.e., a single substitution matrix across all lineages).

When our approach was tested on simulated protein data, it was found that the procedure consistently converges, and that the accuracy of the estimates improves as the number of residues in the sequence alignment grows. We then estimated the amino acid composition of a set of 65 proteins that date to the LUA, and found that tyr, trp, thr, ser, phe, leu, gln, cys and asp occur more frequently, whereas val, ile, his, and glu occur less frequently, in all eight modern species than in the LUA. The frequencies of the remaining amino acids were lower in some and higher in other species relative to the LUA. These inferences are largely consistent with observations made in previous analyses that relied upon very different MP-based methods (Brooks and Fresco, 2002; Brooks et al., 2002). The consensus between the various approaches taken to date indicate those amino acids

for which our findings are most likely to be reliable (see Discussion). Additionally, the inferred amino acid composition of the protein subset in the LUA is more similar to that of modern day thermophilic organisms than to that of mesophilic ones, supporting the theory that the LUA was thermophilic.

Algorithm and application

Estimating ancestral states using ML

Given a sequence alignment and an assumed phylogenetic tree relating these sequences to one another, ML can be used to infer a probability distribution over all amino acids at the root node of the tree; in assigning this distribution, the likelihood of the residues observed at the external nodes of the tree (i.e., in the modern-day sequences) is maximized (Yang et al., 1995). Several assumptions are necessary to make ML inferences of ancestral sequences mathematically tractable. A Markovian process of evolution at each site and independence between sites are assumed. Substitution probabilities for amino acids over the course of evolution are needed; these are typically obtained from empirically derived substitution matrices such as those of Jones et al. (1992). In addition, as noted earlier, a prior distribution of amino acids must be assumed at the root of the tree. (For a full description of ML ancestral sequence reconstruction, see Yang et al. (1995).)

We briefly illustrate the ML method for reconstructing ancestral sequences with an example, at the same time introducing some necessary notation. Assume we are given a set of four aligned extant protein sequences, as well as a tree that represents the phylogenetic relationship of these sequences (Figure 1; only the tree is shown, with

external nodes 1-4, internal nodes 5-7, and internal node 7 as the root node r). The data at site j within these aligned sequences are represented by $\mathbf{x}^{(j)} = \{x_1^{(j)}, x_2^{(j)}, \dots, x_4^{(j)}\}$ where $x_i^{(j)}$ gives the amino acid in the i -th extant sequence at site j . Similarly, $\mathbf{y}^{(j)} = \{y_5^{(j)}, y_6^{(j)}, y_7^{(j)} = y_7^{(j)}\}$ are the amino acids assigned to the internal nodes of the tree, where $y_i^{(j)}$ is the amino acid for the i -th node of the tree at site j . For site j within the aligned sequences, the probability of observing data $\mathbf{x}^{(j)}$ is given by:

$$f(\mathbf{x}^{(j)}; \theta) = \sum_{y_r^{(j)}} p(y_r^{(j)}) f(\mathbf{x}^{(j)} | y_r^{(j)}; \theta) \quad Eq.1$$

where θ refers to the branch lengths of the phylogenetic tree, which we assume to be fixed (these are estimated as described in a subsequent section); $p(i)$ is the prior probability of amino acid i at the root (i.e., given by the ancestral amino acid composition); and $f(\mathbf{x}^{(j)} | y_r^{(j)}; \theta)$ is the conditional probability of observing the data $\mathbf{x}^{(j)}$ given that the reconstruction of the root node at site j is $y_r^{(j)}$. In terms of the tree shown in Figure 1, $f(\mathbf{x}^{(j)}; \theta)$ can be evaluated by summing over all possibilities of $\mathbf{y}^{(j)}$:

$$f(\mathbf{x}^{(j)}; \theta) = \sum_{y_r^{(j)}} \sum_{y_6^{(j)}} \sum_{y_5^{(j)}} p(y_r^{(j)}) P(y_r^{(j)}, y_5^{(j)}, t_5) P(y_r^{(j)}, y_6^{(j)}, t_6) P(y_5^{(j)}, x_1^{(j)}, t_1) \\ P(y_5^{(j)}, x_2^{(j)}, t_2) P(y_6^{(j)}, x_3^{(j)}, t_3) P(y_6^{(j)}, x_4^{(j)}, t_4) \quad Eq.2$$

where $P(u, v, w)$ represents the probability that amino acid v is the observed replacement for amino acid u in time unit w , each time unit coming from the appropriate branch length within the tree.

Using Bayes Law, the probability distribution of residues at the root for site j , $f(y_r^{(j)} | \mathbf{x}^{(j)}; \theta)$, is given by:

$$\frac{p(y_r^{(j)}) f(\mathbf{x}^{(j)} | y_r^{(j)}; \theta)}{f(\mathbf{x}^{(j)}; \theta)} \quad Eq.3$$

Given the assumption of independence between sites, the probability of observing the entire sequence alignment is given by:

$$f(x; \theta) = \prod_j \left(\sum_{y_r^{(j)}} p(y_r^{(j)}) (f(x^{(j)} | y_r^{(j)}; \theta)) \right) \quad \text{Eq.4}$$

and the corresponding log-likelihood score l is:

$$l = \sum_j \log \left(\sum_{y_r^{(j)}} p(y_r^{(j)}) (f(x^{(j)} | y_r^{(j)}; \theta)) \right) \quad \text{Eq.5}$$

For emphasis, we restate that to compute $f(x; \theta)$ or l , $p(i)$ must be assumed. In the following section, we describe how to use EM to find $p(i)$.

Application of EM

EM is a general approach to maximum likelihood parameter estimation in statistical models with hidden variables or missing data (Dempster et al., 1977). Essentially, EM alternates between estimating the parameters and estimating the missing data; however, instead of computing a single value for each missing datum, EM computes a probability distribution. More formally, EM calculates the expected value over the missing data of the log-likelihood score of the observed and missing data as a function of the parameters (E-step), and then re-estimates the parameter values to maximize this expectation (M-step). These steps are repeated until the computed values of the parameters converge. The EM algorithm guarantees that the log-likelihood of the observed data (with the hidden data marginalized) will increase monotonically in each iteration.

Here, the amino acid frequencies of the ancestral sequence at the root, $p(i)$, are the parameters to be estimated, the observed data are the aligned modern-day sequences, and the missing data are the true residues in their ancestral sequence. EM can be applied to estimate the ancestral amino acid composition in the following manner:

- Assume an initial ancestral sequence composition $p(i)$.
- E-step:
 1. For each site j , estimate the probability distribution of residues at that site at the root, $f(y_r^{(j)} | x^{(j)}; \theta)$, using Eq.3 and our current estimate of the ancestral sequence composition $p(i)$.
 2. For each amino acid i , compute its expected number of occurrences, c_i , in the ancestral sequence, by

$$c_i = \sum_j f(y_r^{(j)} = i | x^{(j)}; \theta) \quad \text{Eq.6}$$

- M-step: For each amino acid i , compute the ML estimate of $p(i)$ by dividing the expected number of occurrences of amino acid i in the ancestral sequence, c_i , by the total number of residues in the ancestral sequence, N . This estimate of $p(i)$ is used in the next E-step.

The E-step and M-step are iterated until convergence. In practice, we halt the EM procedure when the change in the log-likelihood score, l , between consecutive iterations does not differ by more than 0.0001. This occurs concomitantly with the stabilization of the estimate of $p(i)$. We thus arrive at a final estimate of $p(i)$. Note that this estimate is not based on an explicit reconstruction of the ancestral sequence.

Application of approach to simulated data

We first tested our approach on simulated sequence data. Because we wished ultimately to apply it to a set of sequences present in the LUA, ancestral sequences were generated with the same amino acid composition as that which we previously estimated for the LUA (Brooks et al., 2002). Six sets of simulated sequence data were generated, starting with ancestral sequences of 300, 1,500, 4,000, 9,000, 15,000, and 20,000 residues in length. For each ancestral sequence, a set of eight descendants was generated based on an assumed phylogenetic tree relating those descendants

[(((1:25,2:25):25,(3:25,4:25):25):25, ((5:25,6:25):25,(7:25,8:25):25):25)] (see Figure 1 for an illustration of how trees are represented in Newick notation) and substitution probabilities from the matrices of Jones et al. (1992). Substitution probabilities were

assumed to be constant at each site in the sequence and along each lineage. Insertions and deletions were not modeled.

When applying our approach to the simulated sequences generated as above, we assumed the same model of evolution (i.e., substitution matrix) as was originally used to generate it. To obtain a phylogenetic tree, the program “protdist” in the phylogenetic software package PHYLIP (Felsenstein, 1993) was used to create a distance matrix for the set of aligned descendant sequences, and an unrooted tree based on this distance matrix was built using PHYLIP’s neighbor-joining (Saitou and Nei, 1987) program. The tree was midpoint rooted.

Application to real sequence data

Sixty-five proteins dating to the LUA were selected as described in detail previously (Brooks et al., 2002). In brief, the selected proteins were required to be present in the three primary lineages (eubacteria, archaea, and eukaryotes), and not to show evidence of horizontal transfer. Consequently, such proteins may be inferred to have a lineage that extends back at least to the LUA. The complete protein set was collected for each descendant species included in the analysis (*Aquifex aeolicus*, *Thermotoga maritima*, *Synechocystis* PCC6803, *Bacillus subtilis*, *Escherichia coli* K12, *Saccharomyces cerevisiae*, *Methanobacterium thermoautotrophicum*, and *Archaeoglobus fulgidus*). Sequences were aligned using the default parameters of ClustalW (Thompson et al., 1994). Alignment positions that contain gaps were deleted, leaving a final alignment of 19,349 residues. As with the simulated data, neighbor-joining was used to infer the tree topology and branch lengths, and the tree was midpoint rooted.

Results

Assessment of method on simulated sequence data

We first assessed the behavior and performance of the method on the sets of simulated data. We investigated whether the approach consistently results in a particular solution, irrespective of the initial $p(i)$ assumed. To this end, the method was first applied to the smallest data set (containing an alignment of length 300), using different initial compositions for the first iteration of nine separate trials. These initial compositions were: 1) that of the known ancestral sequences; 2) that of the descendant sequences; 3) one in which all amino acids were used with equal frequency; 4) an ‘extreme’ composition in which alanine was used with a frequency of 0.9981 and all other amino acids were used with a frequency of 0.0001; and 5-9) five compositions that were randomly generated. The approach consistently reached the same solution regardless of the composition assumed for the first iteration (data not shown). For the remaining datasets, the first three starting compositions given above were tried, and they also led to a single solution. Convergence took between thirteen and twenty-two iterations, depending on the size of the dataset and the initial composition. As expected with EM, the log-likelihood score, l , increased with each iteration until convergence. We next determined how performance of the method changes with alignment length. Measuring average relative error as:

$$1/20 \times \sum_i \frac{|p(i)_{known} - p(i)_{est}|}{p(i)_{known}}$$

we found that the error decreased as the number of residues in the set increased (Figure 2). At 20000 residues, the average relative error was very small (0.023).

Application to the LUA data set

The method was then applied to the set of sixty-five real proteins present in each of eight species for which the last common ancestor is the LUA. Table 1 lists the estimated amino acid composition of this set of proteins in the LUA, the average composition of the set in the modern species, and the composition of the set in each of the eight modern species. Four amino acids, val, ile, glu, and his, are inferred to have decreased in frequency in this sequence set between the LUA and each of the modern species, whereas nine amino acids, phe, tyr, cys, trp, leu, gln, asp, ser, and thr, are inferred to have increased in frequency. The frequencies of the remaining amino acids are greater in some of the modern species and lower in others relative to their inferred frequency in the LUA.

Comparison of LUA amino acid composition with that within modern day mesophiles and thermophiles

We sought to determine whether the amino acid composition inferred for the set of sixty-five proteins in the LUA is more similar to the known composition of the identical set of proteins in either extant mesophiles or extant thermophiles. The Euclidean distance between pairs of vectors of amino acid frequency was used as a measure of similarity. The amino acid composition was calculated for the protein set in three mesophilic (*Helicobacter pylori*, *Mycoplasma pneumoniae*, and *Treponema pallidum*) and three thermophilic (*Aeropyrum pernix*, *Methanococcus jannaschii*, and *Pyrococcus horikoshii*) species. (None of these species was used to estimate the amino acid composition of the LUA.) The Euclidean distance between the amino acid composition observed in each

extant species and that inferred in the LUA was calculated. The mean distance between the amino acid composition observed in these three modern day thermophiles and that inferred for the LUA is 0.001747, with a standard deviation of 0.000719. Comparing compositions observed in the modern day mesophiles with that inferred for the LUA, these values are 0.005056 and 0.001207, respectively. Thus it is evident that the amino acid composition estimated for the LUA is more similar to that observed in modern day thermophiles than in modern day mesophiles, the difference being significant at a P-value of 0.05 (based on a Student's t test).

Discussion

We have presented a method for inferring the amino acid composition of a set of ancestral proteins using ML ancestral sequence reconstruction and EM. On simulated sequence data, the approach leads to identical solutions, starting from several varied amino acid compositions assumed for the initial iteration. These solutions may or may not represent a global optimum: although we started from several randomly generated and one 'extreme' assumed composition, it is theoretically possible, though not likely, that all of these were in the vicinity of a single local maximum. Thus, as with any EM procedure, it is advisable to repeat the method with several different initial starting points. Nevertheless, in practice, the method appears likely to perform well on real sequences when the initial composition assumed is equal to that of the descendants. Moreover, the average relative error in estimates of ancestral amino acid frequencies on simulated data decreases with an increase in the size of the data set, most likely because of the stochastic nature of the underlying evolutionary process. On alignments of 20,000

residues (approximately the length of the LUA protein set considered), the average relative error in estimates is small.

The approach was applied to a set of sixty-five proteins dating to the LUA to estimate the amino acid composition of that set in that ancestor. Comparing the estimated ancestral composition with the observed amino acid composition of the protein set in each of eight modern species, we inferred those amino acids that have increased in frequency between the LUA and today, and those that have decreased. In the past, we have used such inferences to deduce which amino acids were added to the genetic code relatively early and which were added later, assuming that those amino acids that have decreased in frequency over evolution were early additions to the code, and vice versa (Brooks and Fresco, 2002; Brooks et al., 2002). According to this rationale, the findings of the current analysis suggest that val, ile, glu, and his were early, and phe, tyr, cys, trp, leu, gln, asp, ser, and thr were late additions to the genetic code. The data do not permit inferences to be drawn for the remaining amino acids. Four of these inferences differ from those of a previous study (Brooks et al., 2002). In that study, thr, ser and asp were inferred to be early additions, whereas glu was inferred to be a late addition to the genetic code. Because of this discrepancy, we consider these predictions to be questionable. The remaining inferences regarding the relative order of addition of amino acids to the genetic code are consistent with those of our two earlier analyses (Brooks and Fresco, 2002; Brooks et al., 2002). In those studies, two general tendencies were observed. First, amino acids that are believed to have been relatively rare in the prebiotic environment, based on experimental simulations of the prebiotic environment (Miller, 1987) and analysis of the Murchison meteorite (Kvenvolden et al., 1970), most particularly cys, phe,

tyr, and trp, were inferred to be late additions to the code, whereas amino acids believed to have been more abundant, such as val and ile, were inferred to be early additions (Brooks and Fresco, 2002; Brooks et al., 2002). Second, most of the amino acids inferred to be later additions to the code were assigned YNN codons (that is, codons with a pyrimidine in the first codon position), whereas those inferred to be earlier additions were generally assigned RNN codons (Trifonov, 2000; Brooks and Fresco, 2003).

Our findings are also relevant to the ongoing debate over whether the LUA was likely to have been mesophilic or thermophilic (Bocchetta et al., 2000; Brochier and Philippe, 2002). Whereas the inferred moderate G+C content of the small- and large-subunit rRNA of the LUA has been interpreted as evidence for its having been a mesophile (Galtier et al., 1999), the inferred amino acid composition of two proteins in the LUA (using a method distinct from ours) has been interpreted as evidence for its having been a thermophile (DiGiulio, 2001). The findings presented here are consistent with the latter study: we find that the amino acid composition of the set of sixty-five proteins in the LUA inferred using our new method is more similar to the composition observed in extant thermophilic species than in extant mesophilic ones.

In this investigation, we have inferred the amino acid composition of a subset of the proteins in the LUA, not of its entire proteome. The set of ancient, conserved proteins analyzed in the study is composed largely of ribosomal proteins, aminoacyl-tRNA synthetases, polymerases and nucleic acid cleaving enzymes. Due to their need to interact with nucleic acids, these proteins might be expected to use the positively charged amino acids lys and arg more frequently than the proteome as a whole. Although this is in fact observed, the negatively charged amino acids glu and asp also occur more

frequently in the set of 65 proteins than in the whole proteomes (Table 1). These observations are not surprising since these proteins do not generally display skewed isoelectric points. Such differences in amino acid composition between this dataset and the entire proteomes suggest that our analysis might not extend to the ancestral proteome in every detail; nevertheless, the analysis was consistently performed on the same set of proteins, and we expect that the general trends observed are likely to be true.

Furthermore, the 65 ancient, conserved proteins included in our investigation presumably partly retain and reflect the amino acid composition of proteins in the LUA in ways that less ancient proteins in extant proteomes do not, since many of the residues within these proteins have remained unchanged since that early ancestor. Ignoring any bias that may have arisen due to the structural and/or functional requirements of this particular protein set (as discussed above), the modern composition of these proteins should be intermediate between the composition of the set in the LUA and the composition of the modern complete proteomes. Accordingly, it is expected that amino acids that are present at a higher frequency in the set than in modern proteomes have experienced a decrease in frequency since the LUA, whereas those that are present at a lower frequency in the set have experienced an increase. Based on the observed data, this rationale suggests that ala, arg, asp, glu, gly, his, lys, pro, and val have decreased and asn, cys, gln, ile, leu, met, phe, ser, thr, trp, and tyr have increased in frequency since the LUA. This is in agreement with all but two (ile and asp) of the thirteen inferences made in this paper regarding the change in amino acid frequency since the LUA. Thus, comparisons between the modern-day 65 protein dataset and the entire proteomes agree,

at least qualitatively, with the more rigorous analyses performed by our computational method in comparing the same 65 proteins with the inferred composition of its ancestors.

Although we have described a well-justified theoretical approach for estimating the amino acid composition of a set of ancestral sequences through analysis of modern day ones, the estimate made of the amino acid composition of the LUA protein set should not be viewed as definitive. Rather, it represents an attempt, among others we have made, to investigate this problem. Given that estimates resulting from this approach are inherently sensitive to the substitution probabilities assumed, the introduction of, for example, lineage-specific evolutionary models, or rate variation among sites, would be expected to increase their accuracy. We anticipate that such advances will be forthcoming.

Acknowledgements

We wish to thank Carl Kingsford and Elena Nabieva for thoughtful comments on the manuscript, and Sean Eddy for helpful discussions. D.J.B. was supported by predoctoral traineeships from NIH grant 2T32GM07388-22 and NSF grant DGE 9972930 and M.S. was supported by NSF Pecase Grant MCB-0093399 and DARPA grant N66001-02-1-8929. The computational facility utilized for this work was obtained with funds provided to J.R.F. by the Department of Defense through MEDCOM at Fort Detrick, MD.

References

- Bocchetta, M., Gribaldo, S., Sanangelantoni, A. and Cammarano, P. (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**, 366-80.
- Brochier, C. and Philippe, H. (2002) Phylogeny - A non-hyperthermophilic ancestor for bacteria. *Nature*, **417**, 244-244.
- Brooks, D.J., Fresco, J.R., Lesk, A.M. and Singh, M. (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.
- Brooks, D.J. and Fresco, J.R. (2002) Increased frequency of cysteine, tyrosine and phenylalanine residues since the Last Universal Ancestor. *Mol. Cell. Proteomics*, **1**, 125-131.
- Brooks, D.J. and Fresco, J.R. (2003) Greater GNN pattern bias in sequence elements encoding conserved residues of ancient proteins may be an indicator of amino acid composition of early proteins. *Gene*, **303**, 177-85.
- Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington, DC, Vol 5 Suppl 3, pp. 345-352.
- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, **39**, 1-38.
- Di Giulio M. (2001) The universal ancestor was a thermophile or a hyperthermophile. *Gene*, **281**, 11-17.

- Felsenstein, J. 1993 PHYLIP (Phylogeny Inference Package) version 35c Distributed by the author. Department of Genetics, University of Washington, Seattle.
- Galtier, N. and Gouy, M. (1998) Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.*, **15**, 871-879.
- Galtier, N., Tourasse, N. and Gouy, M. (1999) A nonhyperthermophilic common ancestor to extant life forms. *Science*, **283**, 220-221.
- Gerstein, M. (1998) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Des.* **3**, 497-512.
- Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275-282.
- Knight, R.D., Freeland, S.J. and Landweber, L.F. (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* **2**, RESEARCH0010.
- Kreil, D.P. and Ouzounis, C.A. (2001) Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acid Res.* **29**, 1608-1615.
- Kvenvolden, K., Lawless, J., Pering, E., Peterson, E., Flores, J., Ponnampereuma, C., Kaplan, I.R. and Moore, C. (1970) Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite. *Nature*, **228**, 923.
- Miller, S.L. (1987) Which organic compounds could have occurred on the prebiotic earth? *Cold Spring Harbor Symp. Quant. Biol.*, **52**, 17-27.

- Pruess, M., Fleischmann, W., Kanapin, A., Karavidopoulou, Y., Kersey, P., Kriventseva, E., Mittard, V., Mulder, N., Phan, I., Servant, F. and Apweiler, R. (2003) The Proteome Analysis database: a tool for the *in silico* analysis of whole proteomes. *Nucleic Acids Res.*, **31**, 414-417.
- Saitou, N. and Nei, M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406-425.
- Saunders, N.F.W., Thomas, T., Curmi, P.M.G., Mattick, J.S., Kuczek, E., Slade, R., Davis, J., Franzmann, P.D., Boone, D., Rusterholtz, K. et al. (2003) Mechanisms of thermal adaptation revealed from the genomes of the Antarctic *Archaea* *Methanogenium frigidum* and *Methanococcoides burtonii*. *Genome Res.*, **13**, 1580-1588.
- Sueoka, N. (1961) Compositional correlation between deoxyribonucleic acid and protein. *Cold Spring Harbor Symp. Quant. Biol.*, **26**, 35-43.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673-4680.
- Trifonov, E.N. (2000) Consensus temporal order of amino acids and evolution of the triplet code. *Gene*, **261**, 139-151.
- Yang, Z., Kumar, S. and Nei, M. (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, **141**, 1641-1650.

Zhang, Y. and Nei, M. (1997) Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.*, **44** (Suppl 1), S139-S146.

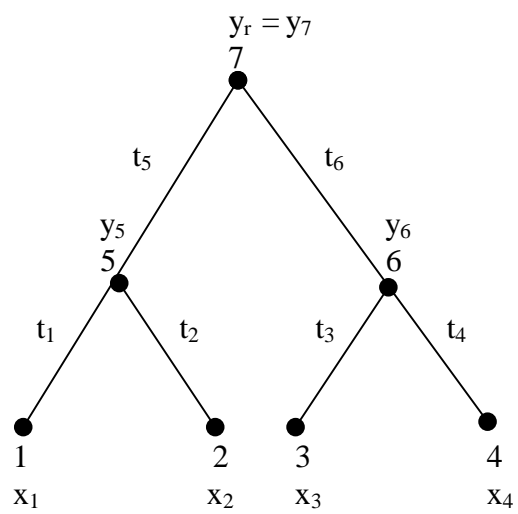


Figure 1

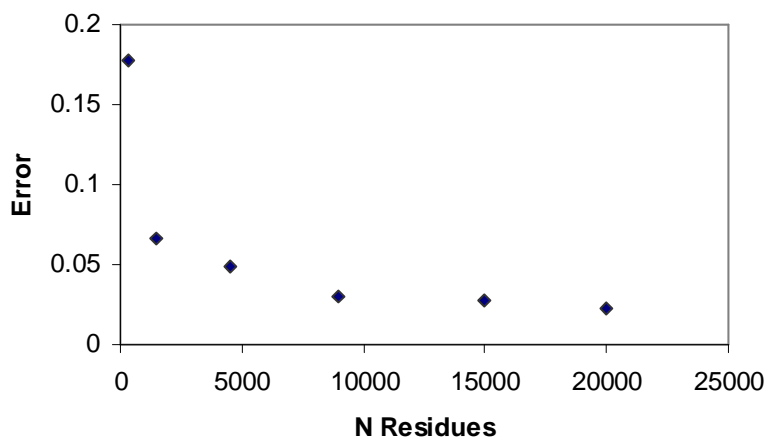


Figure 2

Figure 1. The phylogenetic tree used to illustrate the ML method of ancestral sequence reconstruction. The tree has external nodes 1-4, containing at site j in the aligned sequence data $\mathbf{x}^{(j)} = x_1^{(j)}, x_2^{(j)}, x_3^{(j)}$ and $x_4^{(j)}$, and internal nodes 5, 6 and 7, with associated inferred ancestral states $\mathbf{y}^{(j)} = y_5^{(j)}, y_6^{(j)}$ and $y_7^{(j)} = y_7^{(j)}$. Superscripts are omitted in the figure. Branch lengths between adjacent nodes are represented by t_1 - t_6 . This tree is represented in Newick format as $((x_1:t_1, x_2:t_2):t_5, (x_3:t_3, x_4:t_4):t_6)$. Successively nested parentheses indicate increasingly inclusive clades of taxa. The comma within each set of parentheses separates the taxa to be clustered; the taxon indicator comes first, followed by the length of the branch connecting it to the node that is the last common ancestor of the joined taxa.

Figure 2. Performance of method improves as the number of residues in the ancestral sequence increases. N Residues indicates the number of residues in the ancestral sequence. Error is defined in the text.

Table 1. Amino acid frequencies within a set of sixty-five proteins observed in eight modern species and inferred in the LUA. Species abbreviations are as follows: *Aquifex aeolicus*, Aae; *Thermotoga maritima*, Tma; *Synechocystis* PCC6803, Ssp; *Bacillus subtilis*, Bsu; *Escherichia coli* K12, Eco; *Saccharomyces cerevisiae*, Sce; *Methanobacterium thermoautotrophicum*, Mth; *Archaeoglobus fulgidus*, Afu. The column headed Inferred LUA gives the inferred frequencies within the LUA of the set of sixty-five proteins; that headed Average Modern Set gives the average frequencies in this set within the eight extant species included in the study; that headed Average Modern Proteome gives the average frequencies of the whole proteomes for the eight species.

	Aae	Tma	Ssp	Bsu	Eco	Sce	Mth	Afu	Average Modern Set	Inferred LUA	Average Modern Proteome
Ala	0.0651	0.0660	0.0932	0.0826	0.0990	0.0688	0.0779	0.0814	0.0792	0.0819	0.0726
Arg	0.0571	0.0626	0.0613	0.0519	0.0620	0.0524	0.0744	0.0675	0.0612	0.0685	0.0527
Asn	0.0310	0.0332	0.0324	0.0348	0.0348	0.0429	0.0244	0.0272	0.0326	0.0272	0.0398
Asp	0.0481	0.0491	0.0527	0.0546	0.0564	0.0542	0.0601	0.0518	0.0534	0.0456	0.0516
Cys	0.0077	0.0060	0.0094	0.0043	0.0087	0.0121	0.0104	0.0095	0.0085	0.0040	0.0102
Gln	0.0253	0.0242	0.0503	0.0345	0.0361	0.0368	0.0237	0.0207	0.0315	0.0166	0.0319
Glu	0.0964	0.0930	0.0685	0.0873	0.0699	0.0653	0.0924	0.0944	0.0834	0.1182	0.0764
Gly	0.0705	0.0713	0.0781	0.0751	0.0780	0.0680	0.0745	0.0720	0.0734	0.0733	0.0694
His	0.0191	0.0192	0.0183	0.0202	0.0205	0.0219	0.0227	0.0200	0.0202	0.0237	0.0189
Ile	0.0711	0.0712	0.0635	0.0674	0.0592	0.0680	0.0770	0.0719	0.0687	0.0806	0.0696
Leu	0.0951	0.0907	0.1045	0.0883	0.0899	0.0919	0.0848	0.0857	0.0914	0.0832	0.1010
Lys	0.0957	0.0865	0.0540	0.0745	0.0616	0.0772	0.0581	0.0776	0.0732	0.0874	0.0642
Met	0.0191	0.0227	0.0193	0.0241	0.0282	0.0230	0.0285	0.0234	0.0236	0.0202	0.0246
Phe	0.0393	0.0397	0.0319	0.0340	0.0340	0.0419	0.0362	0.0386	0.0369	0.0308	0.0443
Pro	0.0435	0.0428	0.0463	0.0394	0.0406	0.0411	0.0445	0.0414	0.0425	0.0406	0.0423
Ser	0.0361	0.0417	0.0461	0.0466	0.0438	0.0586	0.0475	0.0413	0.0452	0.0213	0.0614
Thr	0.0449	0.0459	0.0544	0.0550	0.0534	0.0543	0.0450	0.0411	0.0492	0.0390	0.0501
Trp	0.0111	0.0099	0.0106	0.0078	0.0096	0.0111	0.0101	0.0109	0.0101	0.0067	0.0113
Tyr	0.0364	0.0325	0.0255	0.0296	0.0257	0.0329	0.0276	0.0311	0.0301	0.0231	0.0340
Val	0.0873	0.0918	0.0797	0.0880	0.0887	0.0777	0.0801	0.0925	0.0857	0.1080	0.0736