Evolution of Amino Acid Frequencies in Proteins Over Deep Time: Inferred Order of Introduction of Amino Acids into the Genetic Code

Dawn J. Brooks, Jacques R. Fresco,* Arthur M. Lesk,† and Mona Singh‡*

*Department of Molecular Biology, Princeton University †Department of Haematology, Cambridge Institute for Medical Research, United Kingdom; and ‡Department of Computer Science and the Lewis-Sigler Institute for Integrative Genomics, Princeton University

To understand more fully how amino acid composition of proteins has changed over the course of evolution, a method has been developed for estimating the composition of proteins in an ancestral genome. Estimates are based upon the composition of conserved residues in descendant sequences and empirical knowledge of the relative probability of conservation of various amino acids. Simulations are used to model and correct for errors in the estimates. The method was used to infer the amino acid composition of a large protein set in the Last Universal Ancestor (LUA) of all extant species. Relative to the modern protein set, LUA proteins were found to be generally richer in those amino acids that are believed to have been most abundant in the prebiotic environment and poorer in those amino acids that are believed to have been unavailable or scarce. It is proposed that the inferred amino acid composition of proteins in the LUA probably reflects historical events in the establishment of the genetic code.

Introduction

An insight into how proteomic amino acid composition has changed over vast evolutionary time is required for a thorough understanding of the process of protein evolution. Knowledge of the amino acid composition of early proteomes can reveal which of the amino acids have increased and which have decreased in frequency with evolution. Such information could provide clues to the order in which amino acids were introduced into the genetic code and thus into primitive proteins, as suggested previously on the basis of a different approach (Brooks and Fresco 2002).

The recent increase in the number of whole genome sequences has made the analysis of the corresponding inferred proteomes possible. These analyses have included a statistical description of amino acid composition and sequence length (Gerstein 1998*a*), and surveys of both structurally determined (Gerstein 1998*b;* Wolf, Grishin, and Koonin 2000) and predicted folds (Gerstein 1997). Interspecies proteomic comparisons have revealed the phylogenetic distribution of protein families (Tatusov, Koonin, and Lipman 1997); in turn, this knowledge has been used to infer the protein complement of the Last Universal Ancestor (LUA) (Kyrpides, Overbeek, and Ouzounis 1999), the single-celled ancestor that gave rise to the three primary lineages, the eubacteria, archaea, and eukaryotes.

Here, we extend these earlier proteomic characterizations by estimating the amino acid composition of proteins in an ancestral genome. The existing methods used for the reconstruction of ancestral protein sequences, maximum parsimony (MP) and maximum likelihood (ML), are by themselves insufficient to provide accurate estimates of ancestral sequence composition. Where sequences have diverged significantly, MP can only provide partial sequence reconstructions; the identity of residues at many sites in the ancestral sequence will remain ambiguous. Moreover, those sites with assigned residues will be biased toward amino acids that tend to be conserved. On the other hand, ML requires 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 the extant descendants (Yang, Kumar, and Nei 1995). Although Galtier, Tourasse, and Gouy (1999) estimated the GC content of ancestral genomic DNA sequences using an ML approach (not on the basis of sequence reconstruction), a similar approach may not be tractable for protein sequences, given the significantly larger number of parameters to be estimated.

The method we introduce for inferring ancestral amino acid composition is based on the insight that the amino acid composition of conserved residues in present-day proteins, i.e., those residues which are unchanged between an ancestral sequence and any given descendant sequence, is determined by two factors: (1) the amino acid composition within ancestral sequences, and (2) the relative probability of conservation of each amino acid between an ancestral and an extant descendant sequence. Reversing this logic, given the amino acid composition of conserved residues and the relative probability of conservation of each amino acid, the amino acid composition within ancestral sequences can be inferred. Our approaches for identifying and estimating the composition of conserved residues, and estimating the relative probability of conservation of different amino acids are discussed subsequently (see *Methods*).

Simulations of sequence evolution have been employed previously to evaluate alternative methods for reconstructing events in the evolutionary past. For example, they have been used to assess the accuracy of methods for building phylogenetic trees (Saitou and Nei 1987) and reconstructing ancestral protein sequences (Zhang and Nei 1997). We introduce simulations here

Abbreviations: MP, maximum parsimony; ML, maximum likelihood; LUA, Last Universal Ancestor; COG, Clusters of Orthologous Groups.

Key words: amino acid composition, protein evolution, Last Universal Ancestor, substitution probability matrices, genetic code.

Address for correspondence and reprints: Mona Singh, Department of Computer Science, Princeton University, Princeton, New Jersey 08544. E-mail: mona@cs.princeton.edu.

Mol. Biol. Evol. 19(10):1645–1655. 2002

to model errors in the estimate of amino acid composition of ancestral sequences. Using these models, the bias in the estimated ancestral amino acid frequencies might be compensated for, allowing the estimates to be placed within a confidence interval.

Using our approach, we have estimated the amino acid composition of a set of 65 proteins within the LUA. We infer that within this set of proteins many of the amino acids believed to have been most abundant in the prebiotic environment (Miller 1953, 1987; Kvenvolden et al. 1970), including glycine, alanine, aspartic acid, and valine, were used more frequently within proteins of the LUA than within those of modern species. On the other hand, amino acids believed to have been rare or unavailable prebiotically, including cysteine, tryptophan, tyrosine, and phenylalanine, were generally used much less frequently within proteins of the LUA. This may reflect the order of addition of these two groups of amino acids to the genetic code.

Methods

Basis of Approach

The approach is based on the rationale that the amino acid composition of the residues conserved between ancestral and descendant sequences will be determined by the composition of the ancestral sequences and the relative probability of conservation of each amino acid. This relationship may be stated in terms of Bayes' Law:

$$P(i \mid \text{conserved}) = P(i) \times P(\text{conserved} \mid i)/P(\text{conserved}) \quad (1a)$$

where P(i|conserved) denotes the probability of observing amino acid i in the ancestor, given that a residue is conserved between the ancestor and any given descendant, P(i) the probability of amino acid i in the ancestral sequence, P(conserved) the probability that a residue in the ancestor is conserved in any given descendant sequence, and P(conserved|i) the probability of being conserved from the ancestor to any modern descendant, given amino acid i in the ancestral sequence. Our definitions of conserved residues, P(i|conserved), P(i), P(conserved|i), and P(conserved), and their relationship as described by equation (1a) are all illustrated in figure 1*A*. (Residues are defined as conserved on the basis of their identity in modern sequences relative to the ancestral sequence, regardless of their identity at intermediate time points; residues which have changed away from and then back to their initial state are defined as conserved.) Note that according to our definitions of these probabilities, equation (1a) holds for any set of descendant sequences and their true ancestral sequence, regardless of the evolutionary relationship between the sequences and the process by which the descendant sequences evolved.

A rearrangement of equation (1a) shows that the amino acid composition within a set of ancestral sequences may be determined if the remaining probabilities are known:

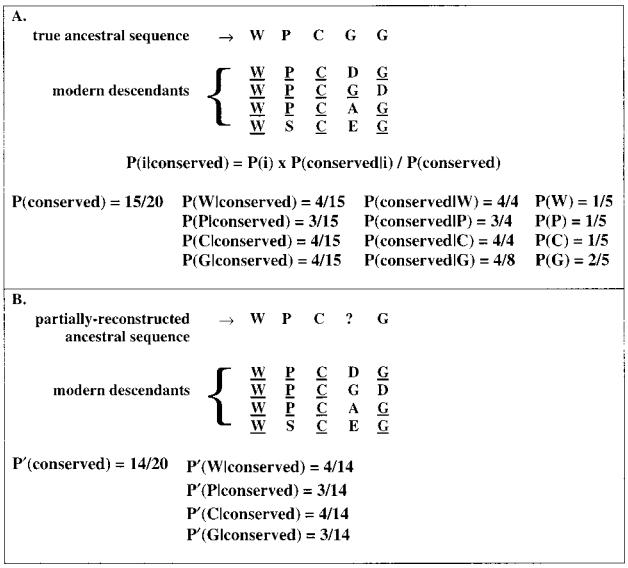

FIG. 1.—*A*, Illustration of the relationship between P(i), P(i|conserved), P(conserved), and P(conserved|i). A simplified schematic showing a section of a true ancestral sequence aligned with its modern descendants. Residues conserved between the ancestor and each descendant are underlined. P(i|conserved), P(conserved), and P(conserved|i) are based on so-identified conserved sequence positions. One can see from this example that the relationship expressed in equation (1b) holds. *B*, Estimation of P(conserved) and P(i|conserved). For estimation of P(conserved) and P(i|conserved), conserved residues are defined as those residues in the modern and inferred ancestral sequences which are identical. (Note that the ancestral sequence is inferred through parsimony and that residues do not have to be identical in all modern sequences to be defined as conserved.) Residues identified as conserved in this manner are underlined. P′(conserved) is the fraction of residues in all the descendant sequences which are identified as conserved, whereas P′(i|conserved) is the fraction of all conserved sites containing residue i. Assuming that the fourth residue in the true ancestral sequence is G, P′(conserved) and P′(G|conserved) are both underestimates, whereas P′(W|conserved), P′(P|conserved), and P′(C|conserved) are overestimates (compare with values in *A*).

$$P(i) = P(i \mid \text{conserved}) \times P(\text{conserved})/P(\text{conserved} \mid i) \quad (1b)$$

When working with real sequences the entire ancestral sequence will not be known, but using estimates for P(i|conserved), P(conserved), and P(conserved|i), we may estimate P(i) (eq. 1b). In the following section, we describe in detail our methods for deriving these estimates and state the assumptions underlying each derivation. In general, the method requires that we assume the following: (1) a Markovian process of evolution, (2) specific amino acid substitution probabilities for one evolutionary time step (e.g., those of Jones, Taylor, and Thornton [1992]), and (3) a molecular clock (equal branch lengths in each lineage). The first two assumptions are common for modeling protein evolution (see, for example, Yang, Kumar, and Nei 1995; Zhang and Nei 1997). Importantly, error in the estimates of P(i), including error arising through the violation of the assumption of a molecular clock, can be modeled and corrected for using simulations. Using statistics describing simulated error in the estimates of P(i), P(i) for a real data set can be placed within a confidence interval.

Estimating Probabilites

Identifying Conserved Residues Using MP Sequence Reconstruction

Let P′(i), P′(conserved), P′(conserved|i), and P′(i|conserved) denote the estimated values of P(i), P(conserved), P(conserved|i), and P(i|conserved), respectively. To estimate P′(conserved) and P′(i|conserved), conserved residues must first be identified. Although conserved residues are frequently defined as residues which are identical in all descendant sequences, this definition is too restrictive for our purposes. Instead we identify, as well as possible, all cases in which a residue remains unchanged between the ancestor and any of its descendant sequences; this is consistent with the definition of P(conserved) in equations (1a) and (1b). To facilitate this process, ancestral sequences are partially reconstructed through MP (Eck and Dayhoff 1966), and these are compared with modern sequences to identify residues conserved between the ancestor and descendant (fig. 1*B*). At each site s along the sequence, the number of identities between the inferred ancestor and each of the descendant sequences is tabulated to determine P′(conserved) at s. For example, if three of the four descendant sequences retain the original residue at site s, P′(conserved) at s is 0.75. P′(conserved) is then averaged over all sites. P′(i|conserved) simply represents the composition of sites identified as conserved between the ancestor and any of the descendants (fig. 1*B*).

As illustrated in figure 1, this method does not identify all instances in which a residue has been conserved between the ancestor and descendant, resulting in errors in P′(conserved) and P′(i|conserved). However, this method is preferable to counting as conserved only those sites identical in all descendants because on the basis of that criterion the actual conserved residues would be much more undercounted and thus P′(conserved) and P′(i|conserved) would be even more biased. Importantly, error in P′(i) arising through error in P′(conserved) and P′(i|conserved) may be modeled through simulation and taken into account in the derivation of the final estimates of P(i) (see *Modeling Error in P′(i) Through Simulation* below).

Estimating P(conserved|i) Using Subsitutition Probability Matrices

In theory, in addition to P′(conserved) and P′(i|conserved), P′(conserved|i) could also be obtained through a comparison of descendant sequences with partially reconstructed ancestral sequences. However, obtaining estimates of all three probabilities in such a manner would lead to values of P′(i) equivalent to the composition of the partially reconstructed sequences (once the values are normalized to sum to one); these P′(i) values are expected to be biased toward those amino acids which have a high relative probability of conservation (this expectation was confirmed through simulations; data not shown). To overcome this difficulty, substitution probability matrices based on empirical protein data are introduced as an independent means for obtaining P′(conserved|i).

A PAM 1 matrix represents the global average substitution probabilities that will result in approximately one substitution per 100 residues when applied to a sequence with the same average composition as the sequence set used to derive the matrices. Substitution probabilities for greater evolutionary distances can be computed as powers of the PAM 1 matrix (Dayhoff, Schwartz, and Orcutt 1978, pp. 345–352); this requires the assumption that substitution probabilities are independent of time and position within the sequence. For modeling the composite evolution of a large sequence set in which only average substitution probabilities are required, this assumption is a reasonable first approximation, one which is frequently used in modeling protein evolution (Zhang and Nei 1997). Values for P(conserved|i) are then represented by the diagonal elements of the PAM probability matrix corresponding to a given evolutionary distance (Dayhoff, Schwartz, and Orcutt 1978). That is, P(conserved|i) is the probability that amino acid i is replaced by itself over the specified evolutionary distance.

Using this approach to estimate P(conserved|i) thus requires that a model of evolution based on a set of empirical matrices be assumed, which is common for the evolutionary analyses of proteins. Because a molecular clock is assumed, regardless of the evolutionary relationships between the descendant sequences, for any amino acid in the ancestor the expected fraction of descendant sites conserved is equal to the probability of conservation for that amino acid at a given evolutionary distance. As noted previously, error in the estimates of P(i) arising through deviation from the assumed molecular clock may be modeled and compensated for through simulation.

The question remains as to how to choose the correct PAM distance for estimating P(conserved|i). Given the diagonal elements $P(\text{conserved}|i)_k$ of a PAM matrix of distance k and P(i) of the ancestral sequences, $P(\text{conserved})_k$, the probability that any residue is conserved at evolutionary distance k, is:

$$P(\text{conserved})_k = \sum (P(i) \times P(\text{conserved}|i)_k) \quad (2a)$$

(Dayhoff, Schwartz, and Orcutt 1978).

Conversely, given P(i) and P(conserved) of a set of sequences, the PAM distance k separating these sequences from their ancestor may be determined by finding the distance k for which:

$$P(\text{conserved}) - \sum (P(i) \times P(\text{conserved}|i)_k) = 0 \quad (2b)$$

An exact solution to equation (2b) may be found using Newton's method (Gonnet and Hallett 2000, pp. 207–210); a simple approximate solution can be obtained by enumerating through integral values of k and choosing the k for which

$$\left|\log\left(P(\text{conserved})/\sum (P(i) \times P(\text{conserved}|i)_k)\right)\right| \quad (2c)$$

is minimized, thereby keeping the ratio of P(conserved) and its estimate $\Sigma(P(i) \times P(\text{conserved}|i)_k)$ as close as possible to 1.0.

In our approach, although P(conserved) and P(i|conserved) may be estimated as described previously, P(i) and the distance k from which P(conserved|i) may be estimated are both unknown. We therefore developed the following method to identify the matrix distance separating ancestral and offspring sequences, and thus P(conserved|i) and P(i), given P(conserved) and P(i|conserved). PAM matrices of successively increasing distance k are used to supply $P(\text{conserved}|i)_k$. The method empirically determines the distance k which best fits our estimates of P(conserved) and P(i|conserved) as follows:

(1) For each k (a) estimate P(i) by $P(i)_k = P(i|\text{conserved}) \times P(\text{conserved})/P(\text{conserved}|i)_k$ and (b) because these probabilities should sum to 1.0, calculate the normalization factor α_k such that $\alpha_k \Sigma P(i)_k = 1.0$.

(2) Choose the matrix for which $|\log \alpha_k|$ is minimized, i.e., the matrix k for which the probability estimates of P(i) require the least amount of normalization.

The corresponding normalized values of $P(i)_k$ indicate P(i). It may be shown that our method finds the integer k whose estimates of P(conserved|i) and P(i) minimize the term given in equation (2c) (see *Appendix*). The same method may be used to derive P′(conserved|i) and P′(i), given P′(conserved) and P′(i|conserved).

Assuming that a representative subset of proteins within a genome is being analyzed, substitution matrices based on good global alignments of generic protein families are most likely to accurately reflect probabilities relating to the evolution of the set. Such matrices include the original PAM matrix of Dayhoff (Dayhoff, Schwartz, and Orcutt 1978) and its updated version by Jones, Taylor, and Thornton (1992). Because the matrices of Gonnet, Cohen, and Benner (1992) are based upon alignments of highly divergent homologs, which are inherently difficult to align, for our purposes the matrices of Jones, Taylor, and Thornton (1992) are preferable. (These matrices are based on alignments of relatively closely related proteins [>85% identity].) The matrices of Jones, Taylor, and Thornton (1992), based on release 22 of the SWISS-PROT databank, the current version of which is described in Bairoch and Apweiler (2000), were therefore used for all the analyses described here. This series is frequently used to represent substitution probabilities for evolutionary modeling (see, for example, Yang, Kumar, and Nei 1995; Zhang and Nei 1997). It should be noted that neither matrices based on local alignments, such as BLOSUM, which are biased toward protein interiors (Henikoff and Henikoff 1992), nor matrices derived from specific protein subclasses, such as transmembrane proteins (e.g., Jones, Taylor, and Thornton 1994), are suitable to provide estimates of P(conserved|i).

Lineage-specific versus pooled estimates

Rather than pooling data from multiple species to arrive at P′(conserved), P′(i|conserved), P′(conserved|i), and ultimately P′(i), independent estimates of P(i) for each lineage can be made and then averaged. This allows the elimination of the assumption of a molecular clock. However, the differences in P′(i) arrived at using either pooled or lineage-specific data were insignificant (data not shown). Because pooling data from all lineages facilitates modeling and correction for error in the estimates, estimates based on pooled data were used.

Estimating Amino Acid Composition in the LUA

Choice of Protein Set

Although there has been much recent discussion of the nature of the LUA (see for example Woese 1998), it is sufficient for our purposes to think of the LUA as either a heterogeneous or a homogeneous population of organisms which eventually diverged into the three primary lineages. On the basis of their presence in eubacteria, archaea, and eukaryotes, approximately 325 proteins are proposed to have been present in the LUA (Kyrpides, Overbeek, and Ouzounis 1999).

Orthologous proteins (homologs which arose through speciation) that are present in a group of modern day species may be inferred to have been present in their last common ancestor. To assemble a set of proteins that were likely present in the LUA, we used the Clusters of Orthologous Groups (COG) database (Tatusov, Koonin, and Lipman 1997; Tatusov et al. 2001) to select all orthologous proteins common to a group of eight species which included members from each of the primary lineages: two archaea, *Archaeoglobus fulgidus* and *Methanobacterium thermoautotrophicum*; one eukaryote, *Saccharomyces cerevisiae*; and five eubacteria, *Aquifex aeolicus*, *Thermotoga maritima*, *Synechocystis* PCC6803, *Escherichia coli*, and *Bacillus subtilis*. These species were chosen because their phylogenetic diversity was expected to maximize the extent and accuracy of ancestral sequence reconstruction. Two hundred fifteen protein families in the COG database are present in all eight species.

Inclusion in the reconstruction of paralogs (homologs arising through gene duplication) which arose before the speciation event separating the species would invalidate the assumed species phylogenetic tree. In the COG database, both orthologs and closely related paralogs are grouped together into protein families. To minimize the possibility of including incorrect paralogs in sequence reconstructions, COG protein families were included either if they had only one protein member per species for all species other than yeast or if the existing paralogs could clearly be determined to have arisen after species divergence. (The tree provided with each COG protein family was used to resolve all issues regarding the protein phylogenetic tree.) For many protein families, yeast has a paralog of mitochondrial origin in addition to the true ortholog. The yeast homolog which clusters with the archaeal species was assumed to represent the true ortholog and was chosen for the analysis. This restriction reduced the protein set to 105 families.

Because lateral transfer also invalidates the assumed species tree, it is important to minimize the number of laterally transferred proteins included in the set. Whole genome analyses have suggested that species

phylogenetic trees are broadly consistent with the small subunit rRNA (SSU rRNA) tree (and the trees of the majority of informational proteins, i.e., proteins playing a role in replication, transcription, or translation [Fitz-Gibbon and House 1999; Teichmann and Mitchison 1999]). Therefore, we eliminated from our set all proteins whose trees were not consistent with the SSU rRNA tree (Olsen, Woese, and Overbeek 1994). For this purpose, we applied the criterion that the yeast and archaeal proteins should cluster on the protein tree. After this restriction was applied, 65 proteins remained in the set. These proteins are primarily classified as informational, most playing a role in translation. The COG protein families included in the analysis are listed in table 1.

Application of Method

Multiple sequence alignments created by the program Clustal W (1.74) (Thompson, Higgins, and Gibson 1994) were input to an MP program, protpars, part of the phylogenetic software package PHYLIP (Felsenstein 1993). The phylogenetic tree based on ss rRNA sequence data (Olsen, Woese, and Overbeek 1994) was assumed for sequence reconstruction. On the basis of the comparison of the reconstructed and extant sequences, conserved residues were identified, allowing $P'(i|\text{conserved})$ and $P'(\text{conserved})$ to be derived. Using these estimates, $P'(\text{conserved}|i)$ and $P'(i)$ were determined as described previously. In total, 164,730 residues from the eight species were analyzed to arrive at $P'(i)$. We refer to this estimate as $P'(i)_{LUA}$. Table 2 shows the values of $P'(i|\text{conserved})$, $P'(\text{conserved})$, and $P'(\text{conserved}|i)$ used to arrive at $P'(i)_{LUA}$.

Modeling Error in P′(i) Through Simulation

Simulations are commonly employed to model sequence evolution; typically they are used to assess the performance of alternative phylogenetic methods on known data (see for example Zhang and Nei 1997). Instead, we employ simulations to model and correct for error in $P'(i)$ so that confidence intervals may be determined for $P(i)_{LUA}$. We anticipate error in $P'(i|\text{conserved})$, $P'(\text{conserved})$, and $P'(\text{conserved}|i)$, and consequently in $P'(i)$ because of the inability to identify all true conserved residues (fig. 1*B*) and violation of the assumption of a molecular clock.

The simulations were performed as follows: a set of n ancestral sequences of determined length was randomly generated according to a given P(i). The resulting ancestral amino acid composition is referred to as $P(i)_S$. The ancestral sequences were evolved to produce descendant sequences based on a phylogenetic tree with unequal branch lengths reflecting average evolutionary distances within the protein set and a substitution probability matrix representing PAM 1 (Jones, Taylor, and Thornton 1992). $P'(i)_S$ of the ancestral sequence set was derived using our method, and $P(i)_S - P'(i)_S$ was determined. This difference is the error in $P'(i)_S$.

The modeled sources of error were systematic and predictable. First, $P'(\text{conserved})$ is an underestimate of P(conserved) (values of 0.3425 and 0.4584, respectively). This underestimate can be explained by the fact that any sequence position in which the residue in at least one of the descendants has been conserved, but for which an ancestral residue was not assigned through MP, will lead to an undercount of P(conserved) (see fig. 1*B*). The underestimate of P(conserved) causes $P'(\text{conserved}|i)_k$ to be chosen from a greater PAM distance k than would the true value of P(conserved) (the lower the conservation between sequences, the greater the evolutionary distance between them; see eq. (2a)).

To understand how bias in $P'(\text{conserved})$ and $P'(\text{conserved}|i)$ interact to result in bias in $P'(i)$, it is useful to consider the ratio $P'(\text{conserved}):P'(\text{conserved}|i)$ (see eq. (1b)). With increasing PAM distance, the ratio $P(\text{conserved}):P(\text{conserved}|i)$ for any amino acid i diverges from unity (fig. 2). This trend holds for PAM = 1 to 150, the range of distances relevant to our data set. For those amino acids for which $P(\text{conserved}):P(\text{conserved}|i) < 1.0$, such as tryptophan, the ratio decreases with increasing PAM, whereas for those amino acids for which $P(\text{conserved}):P(\text{conserved}|i) > 1.0$, such as alanine, the ratio increases (fig. 2). Therefore, as a consequence of the underestimate of P(conserved), it is observed that for those amino acids for which $P(\text{conserved}):P(\text{conserved}|i) < 1.0$, $P'(\text{conserved}):P'(\text{conserved}|i)$ is an underestimate ($P'(\text{conserved}):P'(\text{conserved}|i)/P(\text{conserved}):P(\text{conserved}|i) < 1.0$; table 3, columns 2 and 3; fig. 2). In contrast, for those amino acids for which $P(\text{conserved}):P(\text{conserved}|i) > 1.0$, this ratio is an overestimate (table 3, columns 2 and 3; fig. 2).

On the other hand, $P'(i|\text{conserved})$ is generally an underestimate for those amino acids for which $P(\text{conserved}):P(\text{conserved}|i) > 1.0$ and an overestimate for those for which $P(\text{conserved}):P(\text{conserved}|i) < 1.0$ (table 3, column 4), arginine being the only exception. These observed biases can be explained by the fact that those amino acids with a relatively high $P(\text{conserved}|i)$ are more likely to be conserved in enough descendant sequences such that the ancestral residue can be assigned, and therefore, they are more likely to be correctly identified as conserved than those amino acids with a relatively low $P(\text{conserved}|i)$. Thus, values for $P'(i|\text{conserved})$ are biased in favor of amino acids with a relatively high $P(\text{conserved}|i)$ (table 3, column 4).

Qualitatively, bias in $P'(\text{conserved}):P'(\text{conserved}|i)$ and $P'(i|\text{conserved})$ counteract in equation (1b). For those amino acids with a relatively low $P(\text{conserved}|i)$, $P'(i|\text{conserved})$ is an underestimate, whereas $P'(\text{conserved}):P'(\text{conserved}|i)$ is an overestimate (table 3, columns 2–4). The reverse is true for those amino acids with a relatively high $P(\text{conserved}|i)$. Quantitatively, on the other hand, bias in $P'(i|\text{conserved})$ and $P'(\text{conserved}):P'(\text{conserved}|i)$ do not precisely offset, so there is net error in $P'(i)$.

The predictable nature of the bias revealed in this way provided confidence that simulations could be used to model and correct for error in estimates based on real sequence data.

Table 1
COG Protein Families Included in the LUA Protein Set

COG ID	Protein Name
COG0012	Predicted GTPase
COG0013	Alanyl-tRNA synthetase
COG0016	Phenylalanyl-tRNA synthetase alpha subunit
COG0030	Dimethyladenosine transferase
COG0048	Ribosomal protein S12
COG0049	Ribosomal protein S7
COG0051	Ribosomal protein S10
COG0052	Ribosomal protein S2
COG0060	Isoleucyl-tRNA synthetase
COG0063	Predicted sugar kinase
COG0072	Phenylalanyl-tRNA synthetase beta subunit
COG0080	Ribosomal protein L11
COG0081	Ribosomal protein L1
COG0085	DNA-directed RNA polymerase beta subunit/140 kD subunit
COG0087	Ribosomal protein L3
COG0088	Ribosomal protein L4
COG0089	Ribosomal protein L23
COG0090	Ribosomal protein L2
COG0091	Ribosomal protein L22
COG0092	Ribosomal protein S3
COG0093	Ribosomal protein L14
COG0094	Ribosomal protein L5
COG0096	Ribosomal protein S8
COG0097	Ribosomal protein L6
COG0098	Ribosomal protein S5
COG0099	Ribosomal protein S13
COG0100	Ribosomal protein S11
COG0102	Ribosomal protein L13
COG0103	Ribosomal protein S9
COG0112	Glycine hydroxymethyltransferase
COG0125	Thymidylate kinase
COG0126	3-Phosphoglycerate kinase
COG0143	Methionyl-tRNA synthetase
COG0164	Ribonuclease HII
COG0172	Seryl-tRNA synthetase
COG0177	Predicted EndoIII-related endonuclease
COG0180	Tryptophanyl-tRNA synthetase
COG0184	Ribosomal protein S15P/S13E
COG0185	Ribosomal protein S19
COG0186	Ribosomal protein S17
COG0197	Ribosomal protein L16/L10E
COG0200	Ribosomal protein L15
COG0201	Preprotein translocase subunit SecY
COG0202	DNA-directed RNA polymerase alpha subunit/40 kDa subunit
COG0237	Dephospho-CoA kinase
COG0244	Ribosomal protein L10
COG0250	Transcription antiterminator
COG0256	Ribosomal protein L18
COG0258	5′–3′ exonuclease (including N-terminal domain of PolI)
COG0441	Threonyl-tRNA synthetase
COG0442	Prolyl-tRNA synthetase
COG0452	Phosphopantothenoylcysteine synthetase/decarboxylase
COG0459	Chaperonin GroEL (HSP60 family)
COG0468	RecA/RadA recombinase
COG0495	Leucyl-tRNA synthetase
COG0522	Ribosomal protein S4 and related proteins
COG0525	Valyl-tRNA synthetase
COG0532	Translation initiation factor 2 (GTPase)
COG0533	Metal-dependent proteases with possible chaperone activity
COG0541	Signal recognition particle GTPase
COG0550	Topoisomerase IA
COG0552	Signal recognition particle GTPase
COG0575	CDP-diglyceride synthetase
COG0592	DNA polymerase III beta subunit
COG1758	DNA-directed RNA polymerase subunit K/omega

Table 2
Deriving Estimated Amino Acid Composition in the LUA

Amino acid	$P(i)_{modern}$[a]	P′(i|conserved)[b]	P′(conserved|i)[c]	P′(conserved) P′(conserved|i)[d]	$P'(i)_{LUA}$[e]
ala	0.0777	0.0818	0.2063	1.3001	0.1061
arg	0.0627	0.0689	0.2500	1.0728	0.0737
asn	0.0336	0.0226	0.1733	1.5475	0.0349
asp	0.0542	0.0557	0.2620	1.0238	0.0569
cys	0.0078	0.0045	0.4095	0.6549	0.0030
gln	0.0315	0.0154	0.2244	1.1953	0.0184
glu	0.0859	0.0876	0.2960	0.9060	0.0791
gly	0.0730	0.1214	0.3848	0.6970	0.0844
his	0.0192	0.0201	0.1874	1.4310	0.0287
ile	0.0666	0.0619	0.2212	1.2127	0.0748
leu	0.0891	0.1029	0.3970	0.6755	0.0693
lys	0.0776	0.0676	0.2986	0.8981	0.0606
met	0.0241	0.0141	0.1772	1.5134	0.0212
phe	0.0361	0.0331	0.3914	0.6852	0.0226
pro	0.0435	0.0578	0.3373	0.7953	0.0459
ser	0.0466	0.0296	0.1689	1.5883	0.0469
thr	0.0487	0.0393	0.1771	1.5143	0.0594
trp	0.0102	0.0102	0.6068	0.4420	0.0045
tyr	0.0300	0.0224	0.4085	0.6566	0.0147
val	0.0817	0.0831	0.2338	1.1471	0.0950

[a] $P(i)_{modern}$ is the average amino acid composition for the protein set in the eight modern species.

[b] P′(i|conserved) is the estimated probability of observing amino acid i, given that a residue is conserved.

[c] P′(conserved|i) is the estimated probability that a residue is conserved between an ancestral and descendant sequence, given that it was amino acid i in the ancestral sequence.

[d] P′(conserved) is the estimated probability that a residue is unchanged between the ancestral and descendant sequences. P′(conserved) = 0.2682.

[e] $P'(i)_{LUA}$ is the estimated amino acid composition in the LUA before correction for simulated error (see Table 3).

Placing a Confidence Interval on P(i) in the LUA

Simulations were performed with parameters chosen to model the LUA data set. To represent the ancestral sequence set, 65 proteins of 320 residues each (the average length of the analyzed sequences) were randomly generated using $P'(i)_{LUA}$. To generate a phylogenetic tree representing the sequence set, a distance matrix containing average pairwise distances between the species for the entire LUA protein set was first determined using the program protdist of the PHYLIP package (Felsenstein 1993). The program Fitch of the PHYLIP package was used to convert these distances into an unrooted phylogenetic tree with branch lengths. This tree was converted into a tree with its root between the eubacterial lineage and the lineage giving rise to the archaea and eukaryotes. After simulated evolution to create eight descendant sequences, 166,400 residues (a number similar to the 164,730 residues of the real data set) were available to derive $P'(i|conserved)_S$, $P'(conserved)_S$, $P'(conserved|i)_S$, and thereby $P'(i)_S$. $P(i)_S - P'(i)_S$, or the error in $P'(i)_S$, was then determined for each simulation. One hundred simulations were performed to derive statistics (mean and standard deviation) of the error in $P'(i)_S$. This number of simulations was sufficient for the values of these statistics to stabilize.

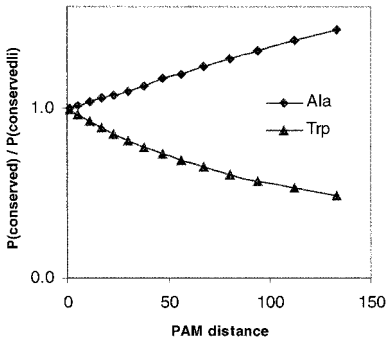

FIG. 2.—P(conserved)/P(conserved|i) diverges from unity with increasing PAM distance. The change in the ratio P(conserved)/P(conserved|i) with change in PAM distance is shown for two representative amino acids: one, alanine, for which P(conserved)/P(conserved|i) > 1.0, and the other, tryptophan, for which P(conserved)/P(conserved|i) < 1.0. The same trend is observed for all other amino acids.

The mean and standard deviation of error in $P'(i)_S$ over 100 simulations (table 4, columns 4 and 5) were used to place $P(i)_{LUA}$ within 95% confidence intervals. First, biases in $P'(i)_{LUA}$ were corrected by adding to them the mean simulated error (table 4, column 6). Lower and upper bounds of 95% confidence intervals for $P(i)_{LUA}$ were determined by subtracting or adding, respectively, twice the standard deviation of error in $P'(i)_S$ to these corrected values (table 4, columns 7 and 8). These confidence intervals are based on the assumption that the simulations accurately model error in $P'(i)_{LUA}$.

Summary of the Method Used for Estimating Composition of Protein Set in LUA

We have explained in detail previously in this article each step of our method and its application to infer

Table 3
Sources of Error in P′(i)[a]

Amino acid	P(conserved) / P(conserved|i)	P′(conserved)/P′(conserved|i) / P(conserved)/P(conserved|i)	P′(i|conserved) / P(i|conserved)
ser	1.3668	1.1065	0.7912
thr	1.2903	1.1008	0.8354
asn	1.2652	1.1173	0.7842
met	1.2113	1.1193	0.7339
ala	1.1833	1.0657	0.9100
his	1.1673	1.1051	0.8241
ile	1.1587	1.0382	0.9208
val	1.1239	1.0246	0.9501
gln	1.0532	1.0726	0.9288
arg	1.0211	1.0292	1.0229
asp	1.0154	1.0072	0.9573
glu	0.9385	0.9836	1.0300
lys	0.9214	0.9860	1.0178
pro	0.8332	0.9678	1.0905
leu	0.7753	0.9272	1.1947
gly	0.7750	0.9404	1.1222
phe	0.7722	0.9309	1.1126
tyr	0.7465	0.9292	1.1357
cys	0.7436	0.9217	1.0724
trp	0.5988	0.8423	1.2143

[a] All probabilities represent mean values derived from 100 simulations.

Table 4
Change in Amino Acid Composition Between LUA and Modern Genomes

Amino acid	$P(i)_{modern}$[a]	$P'(i)_{LUA}$[b]	Mean Error[c]	S Error[c]	Corrected $P'(i)_{LUA}$[d]	95% Confidence Interval on P(i) in LUA[e]		Lower $P(i)_{LUA}/P(i)_{modern}$	Upper $P(i)_{LUA}/P(i)_{modern}$
						Lower Bound	Upper Bound		
ala	0.0777	0.1061	0.0048	0.0027	0.1109	0.1055	0.1163	1.358	1.497
arg	0.0627	0.0737	−0.0087	0.0022	0.0650	0.0606	0.0694	0.967	1.107
asn	0.0336	0.0349	0.0048	0.0019	0.0397	0.0359	0.0435	1.070	1.296
asp	0.0542	0.0569	0.0008	0.0017	0.0577	0.0543	0.0611	1.001	1.127
cys	0.0078	0.0030	0.0000	0.0003	0.0030	0.0024	0.0036	0.306	0.459
gln	0.0315	0.0184	−0.0004	0.0014	0.0180	0.0152	0.0208	0.482	0.659
glu	0.0859	0.0791	−0.0006	0.0023	0.0785	0.0739	0.0831	0.860	0.967
gly	0.0730	0.0844	−0.0063	0.0018	0.0781	0.0745	0.0817	1.020	1.119
his	0.0192	0.0287	0.0034	0.0016	0.0321	0.0289	0.0353	1.503	1.836
ile	0.0666	0.0748	0.0010	0.0025	0.0758	0.0708	0.0808	1.064	1.214
leu	0.0891	0.0693	−0.0120	0.0015	0.0573	0.0543	0.0603	0.609	0.677
lys	0.0776	0.0606	0.0014	0.0019	0.0620	0.0582	0.0658	0.750	0.847
met	0.0241	0.0212	0.0054	0.0014	0.0266	0.0238	0.0294	0.989	1.221
phe	0.0361	0.0226	−0.0013	0.0010	0.0213	0.0193	0.0233	0.534	0.645
pro	0.0435	0.0459	−0.0024	0.0015	0.0435	0.0405	0.0465	0.931	1.0369
ser	0.0466	0.0469	0.0054	0.0026	0.0523	0.0471	0.0575	1.012	1.235
thr	0.0487	0.0594	0.0030	0.0021	0.0624	0.0582	0.0666	1.196	1.368
trp	0.0102	0.0045	−0.0003	0.0003	0.0042	0.0036	0.0048	0.354	0.472
tyr	0.0300	0.0147	−0.0008	0.0007	0.0139	0.0125	0.0153	0.417	0.510
val	0.0817	0.0950	0.0026	0.0028	0.0976	0.0920	0.1032	1.127	1.264

[a] $P(i)_{modern}$ is the average amino acid composition for the protein set in the eight modern species.
[b] $P'(i)_{LUA}$ is the estimated amino acid composition in the LUA (see Table 2) before correction for simulated error.
[c] Mean and standard deviation (s) error are statistics describing P(i)s − P′(i)s derived from 100 simulations, as described in the text.
[d] Corrected $P'(i)_{LUA} = P'(i)_{LUA}$ + mean error.
[e] The lower and upper bounds, respectively, of the 95% confidence interval on $P(i)_{LUA}$ (equal to corrected $P'(i)_{LUA} \pm 2$ s error) are given.

the amino acid composition of a set of proteins in the LUA. It may be helpful to the reader if we now summarize it. First, the protein set was chosen. Next, the proteins were aligned and partial ancestral sequences inferred. Conserved sequence positions were then identified, allowing P(i|conserved) and P(conserved) to be estimated. P(conserved|i) was estimated using a substitution probability matrix. An initial estimate of P(i) could then be made. Finally, sequence simulations were used to model error in estimates of P(i), allowing a confidence interval to be placed on P(i) in the LUA.

Results

Inferred Change in Amino Acid Composition Since the LUA

Using the probabilistic approach and simulations described here, and assuming the Jones, Taylor, and Thornton (1992) series of matrices as a model of amino acid substitution probabilities, the amino acid composition of 65 proteins in the LUA was inferred. Figure 3 shows the change in frequency of individual amino acids within this protein set between the LUA and today (see also table 4). The modern composition upon which this comparison is based represents the average composition of this protein set in the eight modern species examined. Nine amino acids, alanine, asparagine, aspartatic acid, glycine, histidine, isoleucine, serine, theonine, and valine, are inferred to have occurred more frequently in this protein set in the LUA than today (fig. 3*A*). On the other hand, eight amino acids, cysteine, glutamine, glutamic acid, leucine, lysine, phenylalanine, tryptophan, and tyrosine, are inferred to have been used less frequently in the LUA (fig. 3*B*). The confidence intervals on P(arginine), P(methionine), and P(proline) (fig. 3*C*) do not allow an inference to be made as to whether or how these amino acids have changed in frequency since the LUA.

Discussion

Inferred Order of Entry of Amino Acids into the Genetic Code

On the basis of the change in frequency of amino acids between the LUA and today, we may make inferences regarding the establishment of the genetic code (Brooks and Fresco 2002). It is reasonable to assume that as the genetic code evolved, newly assigned amino acids adopted codons used infrequently in coding sequences to minimize the structural disruption of the encoded protein (Osawa et al. 1992). Consequently, new amino acids would have been introduced gradually into existing primitive proteins. Thus, at the time the genetic code became fully established, those amino acids which had been added relatively early would have been overrepresented and those which had been added relatively late would have been underrepresented, relative to the composition of modern proteins. Starting from such early biased amino acid composition, primitive proteins would have proceeded to evolve toward their modern-day compositions. In such a scenario, the amino acids that were introduced into the genetic code relatively ear-

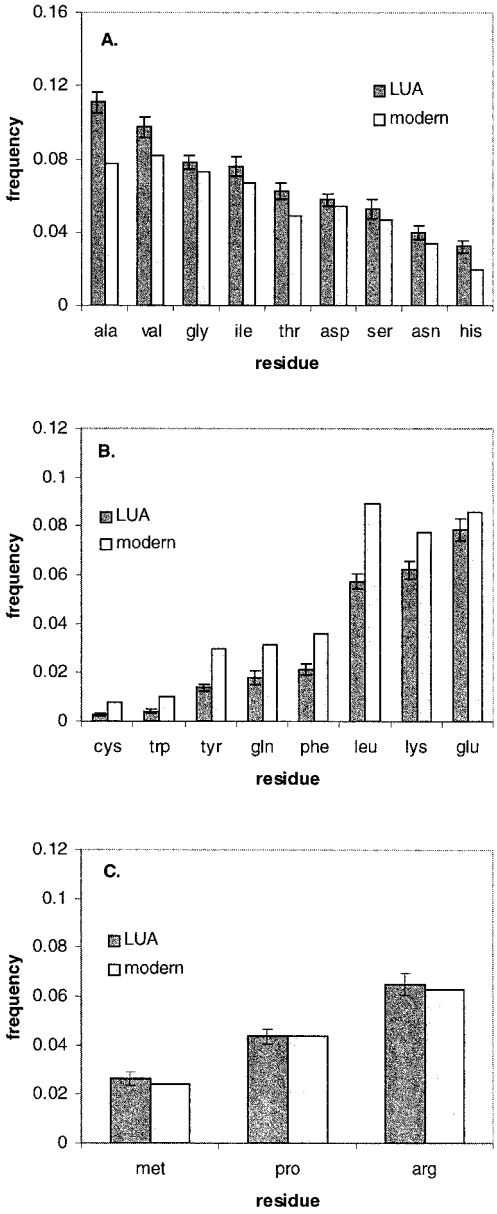

FIG. 3.—Change in amino acid frequencies between the LUA and today. The 95% confidence intervals on frequencies in the LUA given in table 3 are represented here as error bars. ''Modern'' reflects the average amino acid composition for the protein set in the eight modern species. Amino acids with inferred (*A*) decrease, (*B*) increase, or (*C*) no change in frequency between the LUA and today.

ly should have decreased in frequency over the course of evolution, whereas those amino acids added relatively late should have increased in frequency (i.e., between the establishment of the genetic code, the LUA, and today).

The nine amino acids which have decreased in frequency between the LUA and today (fig. 3*A*) may thus be inferred to have been introduced into the code early. Most of these amino acids are among those presumed to have been most abundant in the prebiotic environment, as inferred through spark tube simulations (Miller 1953, 1987) and analysis of the Murchison meteorite (Kvenvolden et al. 1970). In contrast, the eight amino acids which have increased in frequency between the LUA and today (fig. 3*B*), and which are thus inferred to have been late additions to the code, include several of the most biosynthetically complex amino acids (for example, all three aromatic amino acids, which share a common, complex metabolic intermediate, are inferred to have been late additions); most of these are presumed either to have been nonexistent or of very low abundance in the prebiotic environment. Two of these, cysteine and tryptophan, are both conservatively estimated to have been less than half as frequent within this protein set in the LUA than today.

We emphasize that the validity of the inferences drawn in this study depend upon the reliability of the Jones, Taylor, and Thornton (1992) substitution probabilities for modeling evolution over very long time periods and along all lineages. With the development of lineage-specific models of evolution, estimates of ancestral amino acid composition can be expected to improve. In the meantime, although there are undoubtedly limitations to using the matrices of Jones, Taylor, and Thornton (1992) to model evolution since the LUA, we feel they provide the best available estimates of these substitution probabilities. It is noteworthy that previously, on the basis of an independent approach, cysteine, tyrosine, and phenylalanine were inferred to have been used less frequently in proteins of the LUA than today (Brooks and Fresco 2002).

The inferences drawn here regarding the relatively early or late introduction of amino acids into the genetic code are generally consistent with earlier proposals which were based on the presumed presence or absence of various amino acids in the primordial environment (see for example Wong 1975). However, for a few amino acids our assignment as early or late is not in keeping with earlier ideas. For example, histidine and asparagine, believed to have been absent in the prebiotic environment, are both inferred through the present work to have been added to the code relatively early, whereas glutamate, believed to have been present in the prebiotic environment, is inferred to have been added late (figs. 3 and 4). Interestingly, each of these three amino acids share a block of four codons with a second amino acid: histidine with glutamine, asparagine with lysine, and glutamate with aspartate (fig. 4). Codon capture, in which one amino acid loses some of its codons to another, is commonly proposed as a mechanism for introducing amino acids, especially later arriving ones, into the code (Crick 1968; Wong 1975). Consistent with codon capture, it is plausible that one amino acid was added to the four-codon block first and that this amino acid later gave up two of its codons to the second amino acid which now shares the block.

Accordingly, those amino acids which were originally assigned to the codon block (i.e., aspartate, asparagine, and histidine) would have the appearance of being added to the code early, whereas those which were added to the block later through codon capture (i.e., glutamate, lysine, and glutamine) would have the appearance of being added late. Therefore, early and late amino acids do not correspond to a strict chronological order

Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

FIG. 4.—Early or late addition to the genetic code versus presence in the prebiotic environment. Amino acids and their corresponding codons are coded as follows: underlined = early; outline = late; normal = undetermined; and gray highlight = most abundant prebiotic. Assignment of amino acids as early, late, or undetermined is based on this work. Assignment of amino acids as most abundant prebiotic is based on Miller (1987).

of introduction into the code. Instead, as defined here on the basis of the changing amino acid frequencies, early amino acids are probably those which at some point lost some of their codons through codon capture and consequently became less frequent over time within proteins, whereas late amino acids are those which entered the code through codon capture, did not subsequently lose any of their codons, and therefore became more frequent over time (fig. 4). Finally, it is worth noting that the distinction between amino acids inferred here to have been added to the genetic code early or late does not at all correlate with the two main structural classes of the aminoacyl-tRNA synthetases. This is consistent with the earlier suggestion that these enzymes probably had no specific role in the evolution of the genetic code (Woese 2000).

Existing ideas regarding the origin and evolution of the genetic code have been based largely on theoretical investigations and experiments involving oligonucleotide aptamer binding of amino acids (reviewed in Knight, Freeland, and Landweber 1999). The present findings suggest that notwithstanding the impact of mutation over the long course of molecular evolution, with the aid of the appropriate analytical tools and insights, the sequences of contemporary proteins also provide an important avenue for exploring these early evolutionary events.

Acknowledgments

We are grateful to Warren Ewens for providing advice on the statistical treatment of error in the estimate of P(i) and for comments on the manuscript. We also wish to thank Nick Goldman, Robert Osada and Carl Kingsford for thoughtful suggestions on the manuscript. D.J.B. was supported by predoctoral traineeships from NIH grant 2T32GM07388-22 and NSF grant DGE 9972930, A.M.L. is supported by the Wellcome Trust, and M.S. is supported by NSF Pecase Grant MCB-0093399. 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.

APPENDIX

Given P(i) and P(conserved), one may find the matrix of distance k to minimize (2c).

However, if P(i) is unknown, but P(conserved) and P(i|conserved) are known, one can substitute for P(i) as follows:

$$P(i)_k = P(i \mid \text{conserved}) \times P(\text{conserved})/P(\text{conserved} \mid i)_k \quad (3)$$

where $P(i)_k$ is estimated by using $P(\text{conserved}|i)_k$ from the matrix of distance k. Because the $P(i)_k$ are probabilities, the sum over all amino acids is constrained to be one. Let α_k be the normalization factor, that is:

$$\alpha_k \sum P(i)_k = 1.0 \quad (4)$$

and $\alpha_k P(i)_k$ provide estimates of P(i) that may be used in (2c):

$$\left| \log\left(\frac{P(\text{conserved})}{\sum [\alpha_k P(i)_k \times P(\text{conserved} \mid i)_k]} \right) \right|. \quad (5)$$

Substituting for $P(i)_k$ using equation (3) and simplifying gives:

$$\left| \log\left(\frac{P(\text{conserved})}{\alpha_k \sum [P(i \mid \text{conserved}) \times P(\text{conserved})]} \right) \right| \quad (6)$$

$$= \left| \log\left(\frac{P(\text{conserved})}{\alpha_k P(\text{conserved})} \right) \right| \quad (7)$$

$$= \left| \log\left(\frac{1}{\alpha_k} \right) \right| = |\log \alpha_k| \quad (8)$$

Thus, by choosing the value k which minimizes $|\log \alpha_k|$, we are finding the PAM distance k whose corresponding estimates of P(i) and P(conserved|i) minimize (2c).

LITERATURE CITED

BAIROCH, A., and R. APWEILER. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. **28**:45–48.

BROOKS, D. J., and J. R. FRESCO. 2002. Increased frequency of cysteine, tyrosine and phenylalanine residues since the Last Universal Ancestor. Mol. Cell. Proteomics. **1**:125–131.

CRICK, F. H. C. 1968. The origin of the genetic code. J. Mol. Biol. **38**:367–379.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. Atlas of protein sequence and structure. Vol. 5 [Suppl. 3]. National Biomedical Research Foundation, Washington, D.C.

ECK, R. V., and M. O. DAYHOFF. 1966. Atlas of protein sequence and structure. National Biomedical Research Foundation, Silver Spring, Md.

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

FITZ-GIBBON, S. T., and C. H. HOUSE. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. **27**:4218–4222.

GALTIER, N., N. TOURASSE, and M. GOUY. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science **283**:220–221.

GERSTEIN, M. 1997. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. J. Mol. Biol. **274**:562–576.

———. 1998*a*. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. Fold. Des. **3**:497–512.

———. 1998*b*. Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. Proteins **33**:518–534.

GONNET, G. H., M. A. COHEN, and S. A. BENNER. 1992. Exhaustive matching of the entire protein sequence database. Science **256**:1443–1445.

GONNET, G. H., and M. HALLETT. 2000. The Darwin manual. ETH Zurich.

HENIKOFF, S., and J. HENIKOFF. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**:10915–10919.

JONES, D. T., W. M. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**:275–282.

———. 1994. A mutation data matrix for transmembrane proteins. FEBS Lett. **339**:269–275.

KNIGHT, R. D., S. J. FREELAND, and L. F. LANDWEBER. 1999. Selection, history and chemistry: the three faces of the genetic code. Trends Biochem. Sci. **24**:421–427.

KVENVOLDEN, K., J. LAWLESS, E. PERING, E. PETERSON, J. FLORES, C. PONNAMPERUMA, I. R. KAPLAN, and C. MOORE. 1970. Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite. Nature **228**:923.

KYRPIDES, N., R. OVERBEEK, and C. OUZOUNIS. 1999. Universal protein families and the functional content of the last universal common ancestor. J. Mol. Evol. **49**:413–423.

MILLER, S. L. 1953. Production of amino acids under possible primitive earth conditions. Science **117**:528–529.

———. 1987. Which organic compounds could have occurred on the prebiotic earth? Cold Spring Harbor Symp. Quant. Biol. **52**:17–27.

OLSEN, G. J., C. R. WOESE, and R. OVERBEEK. 1994. The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. **176**:1–6.

OSAWA, S., T. H. JUKES, K. WATANABE, and A. MUTO. 1992. Recent evidence for evolution of the genetic code. Microbiol. Rev. **56**:229–264.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

TATUSOV, R. L., E. V. KOONIN, and D. J. LIPMAN. 1997. A genomic perspective on protein families. Science **278**:631–637.

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.

TEICHMANN, S. A., and G. MITCHISON. 1999. Is there a phylogenetic signal in prokaryote proteins? J. Mol. Evol. **49**:98–107.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 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.

WOESE, C. R. 1998. The universal ancestor. Proc. Natl. Acad. Sci. USA **95**:6854–6859.

———. 2000. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol. Mol. Biol. Rev. **64**:202–236.

WOLF, Y. I., N. V. GRISHIN, and E. V. KOONIN. 2000. Estimating the number of protein folds and families from complete genome data. J. Mol. Biol. **16**:897–905.

WONG, J. T. 1975. A co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. USA **72**:1909–1912.

YANG, Z., S. KUMAR, and M. NEI. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641–1650.

ZHANG, Y., and M. NEI. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J. Mol. Evol. **44**(1):S139–S146.

WILLIAM TAYLOR, reviewing editor

Accepted April 26, 2002