Increased Frequency of Cysteine, Tyrosine, and Phenylalanine Residues Since the Last Universal Ancestor*

Dawn J. Brooks‡ and Jacques R. Fresco§

Analysis of extant proteomes has the potential of revealing how amino acid frequencies within proteins have evolved over biological time. Evidence is presented here that cysteine, tyrosine, and phenylalanine residues have substantially increased in frequency since the three primary lineages diverged more than three billion years ago. This inference was derived from a comparison of amino acid frequencies within conserved and non-conserved residues of a set of proteins dating to the last universal ancestor in the face of empirical knowledge of the relative mutability of these amino acids. The under-representation of these amino acids within last universal ancestor proteins relative to their modern descendants suggests their late introduction into the genetic code. Thus, it appears that extant ancient proteins contain evidence pertaining to early events in the formation of biological systems. ***Molecular & Cellular Proteomics 1:125–131, 2002.***

By remaining unchanged over the long course of molecular evolution, conserved residues of ancient proteins might possess significant information regarding early ancestral proteins. We sought to determine whether amino acid frequencies within conserved positions of proteins dating to the last universal ancestor (LUA)[1] of all life indicate that any of the 20 amino acids occurred more or less frequently within early proteins than within their modern descendants. In part, we were motivated by the idea that the amino acid composition of proteins within the LUA might have reflected the order of addition of amino acids to the genetic code, *i.e.* that compared to modern proteins, the composition was relatively richer in amino acids added to the code early and poorer in those added late. Our approach is based on the insight that the amino acid composition of conserved residues of modern-day proteins has been determined by two factors, the composition of the ancestral proteins that gave rise to the extant proteins and the relative mutability of the various amino acids over the course of evolution of the sequences. Therefore, based solely on knowledge of the composition of conserved (*i.e.* unchanged) residues of extant sequences and the relative mutability of each amino acid, it may be possible to make inferences regarding the composition of early ancestral proteins.

The mutability of each amino acid has been determined empirically through pair-wise comparison of aligned homologous protein sequences; mutability is defined as the number of times an amino acid differs at analogous sites of two aligned sequences divided by the total occurrence of that amino acid within the pair of sequences (1). Thus, an amino acid that has mutated relatively frequently over the course of evolution is assigned a high mutability, whereas an amino acid that has mutated relatively infrequently is assigned a low mutability. Amino acids differ in mutability according to the ease with which each particular amino acid may be structurally or functionally replaced by any other within proteins. This depends on the size, shape, hydrophobicity, and charge of each amino acid side chain and its ability to form various types of weak bonds, as well as the structure of the genetic code.

Our approach is based upon the following premise. An amino acid with relatively low mutability is by definition less likely to change over the course of sequence evolution than other amino acids. Therefore, as an original set of ancestral sequences gives rise to successive generations of descendants, the frequency of such an amino acid within conserved positions of those descendants (*i.e.* residues that are unchanged between ancestral and descendant sequences) will increase relative to its frequency within the entire ancestral sequence set. Consequently, the frequency of an amino acid with low mutability within conserved sequence positions of descendant sequences provides an upper limit on its frequency within the ancestral sequences, *i.e.* it must have occurred with a lower frequency within the ancestral sequences as a whole than within the conserved positions of descendant sequences. On the other hand, the frequency of an amino acid with relatively high mutability will decrease over evolution within conserved positions of descendant sequences relative to the entire ancestral sequence set; thus, its frequency within conserved positions provides a lower limit on its frequency within the ancestral sequences. It is important to recognize that these inferences regarding the upper and lower limits of amino acid frequencies within ancestral sequences are com-

From the Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544

Received, August 1, 2001, and in revised form, November 2, 2001
Published, MCP Papers in Press, November 13, 2001, DOI 10.1074/mcp.M100001-MCP200

[1] The abbreviations used are: LUA, last universal ancestor; COG, clusters of orthologous groups.

TABLE I
Summary of rationale for inferring change in frequency of an amino acid over the course of evolution

Mutability	Frequency in conserved / Frequency in non-conserved	Change in frequency over evolution
Low	<1	Increased
Low	>1	?
High	>1	Decreased
High	<1	?

pletely independent of substitution events occurring within non-conserved sequence positions.

As a consequence of the limits specified above, two general types of observations (Table I) would suggest that a change in frequency of an amino acid over evolution within a set of proteins had occurred; if an amino acid with low mutability occurs less frequently within conserved than within non-conserved residues of the extant protein set, its frequency must have increased over evolution, because its frequency within ancestral sequences can be inferred to have been lower than that within conserved residues. Conversely, if an amino acid with high mutability occurs with greater frequency within conserved than non-conserved residues, its frequency can be inferred to have decreased over evolution, because its frequency within ancestral sequences can be inferred to have been higher than that within conserved residues. It is worth remarking that, based on this approach, no inferences regarding changing amino acid frequencies may be made in cases in which an amino acid with low mutability occurs more frequently, or an amino acid with high mutability occurs less frequently, within conserved than non-conserved residues. Nonetheless, this approach may identify some amino acids that have changed in frequency over deep evolutionary time and thereby provide novel insights regarding early proteins. Guided by this rationale, we determined the frequency of each amino acid in conserved and non-conserved sequence elements of a set of extant proteins dating to the LUA in 26 species spanning the three primary lineages.

EXPERIMENTAL PROCEDURES

Choice of Protein Set—Although the nature of the LUA has been the subject of debate (2), for the present work it is sufficient that the LUA was an hetero- or homogeneous population that diverged to form the three primary lineages. Consistent with this view, a set of proteins was selected that can be inferred to have been present within the LUA. The clusters of orthologous groups (COG) database (3), which groups proteins into families based on pairwise comparisons of the protein complements of fully sequenced genomes, was used to assist in the choice of proteins to include in the analysis. Twenty-six major lineages (19 eubacteria, six archaea and one eukaryote) are represented in the COG database. Not all species contribute members to all families in the database; on the other hand, some species contribute more than one member to a particular family.

Our first requirement was that a member of a protein family be present in at least one species of each of the three primary lineages, because this criterion is used to infer that an ancestor of that family was present in the LUA (4). In fact, we required that for any protein family to be included in the study, at least one member had to be present in all 26 species selected from the COG database (for the list of species, see the legend to Table V). This made it possible to assemble a set containing members from the same protein families for each of these species. Although only one eukaryote, *Saccharomyces cerevisiae*, was included in the analysis, this did not in any way limit the ability to identify conserved sequence positions within the protein set or to draw conclusions based on the data obtained. In fact, the very wide phylogenetic representation of both eubacteria and archaea was more than sufficient to identify conserved residues, allowing inferences to be drawn regarding the frequency of certain amino acids within ancestral sequences in the LUA.

The inclusion of proteins that have been laterally transferred between the eubacterial and the archaeal/eukaryotic lineages would confound our ability to identify residues conserved since the LUA. The protein set was therefore chosen so as to minimize inclusion of laterally transferred proteins. The phylogenetic grouping of the archaea and eukaryotes within a lineage distinct from that of the eubacteria, originally based upon the small subunit rRNA tree (5), has been supported by whole genome analysis (6). Therefore, for any protein family to be included in the analysis, it was required that the family member from the one eukaryotic species, *S. cerevisiae*, and the members from the six archaeal species form a cluster that is separate from the members contributed by the eubacterial species on the phylogenetic tree provided with each COG (suggesting that proteins within this family have not been laterally transferred between the eubacterial and the archaeal/eukaryotic lineages). Finally, for the purpose of sequence reconstruction (see below), it had to be assumed that species and protein trees are congruent, an assumption potentially violated by inclusion of paralogs (homologs arising through gene duplication) that arose prior to speciation. Therefore, for inclusion of any COG family in the analysis, it had to have one homolog within each species, whether an ortholog or paralog, that did not invalidate the assumption of species and protein tree congruence.

After these requirements were fulfilled, our protein set consisted of 59 COG families (Table II). Forty-five of these proteins play some role in translation (many are ribosomal proteins), and another seven play a role in transcription, replication, or DNA repair. These all are classified as informational proteins (7), because they function in replication, transcription, or translation. The remaining seven proteins are classified as operational proteins (7), which perform metabolic and other housekeeping roles within the cell. Informational proteins have been found to be less likely to be laterally transferred than operational proteins (7), and because one of the goals in choosing the set was to avoid laterally transferred proteins, the high proportion of informational proteins in the set was both expected and reassuring.

Identification of Conserved Residues—The next step was to identify residues within the 59 proteins from each of the 26 species that have been conserved since the LUA. Sequences were aligned using ClustalW (8). Two approaches were then used to identify conserved residues within each of the descendant sequences. The first was to identify positions in which the amino acid residues in all 26 descendants are identical. We refer to such positions as "identical sites" to distinguish them from conserved residues identified using the second method described below. Identical sites are rare (~2% of sequence sites) and exclude many residues actually conserved between an ancestral sequence and any given descendant sequence.

To identify conserved residues more accurately, maximum parsimony (9) was used to partially reconstruct the ancestral protein sequences in the LUA that gave rise to each family of aligned descendants. The protein parsimony software "protpars" included in the PHYLIP phylogenetic package (10) was used to partially reconstruct ancestral sequences, assuming the phylogenetic tree indicated by small subunit rRNA data (5). Using the inferred ancestral sequence,

TABLE II
COG protein families included in the LUA protein set

COG0013 alanyl-tRNA synthetase, COG0030 dimethyladenosine transferase, COG0060 isoleucyl-tRNA synthetase, COG0495 leucyl-tRNA synthetase, COG0143 methionyl-tRNA synthetase, COG0016 phenylalanyl-tRNA synthetase α-subunit, COG0072 phenylalanyl-tRNA synthetase β-subunit, COG0442 prolyl-tRNA synthetase, COG0081 ribosomal protein L1, COG0244 ribosomal protein L10, COG0080 ribosomal protein L11, COG0102 ribosomal protein L13, COG0093 ribosomal protein L14, COG0200 ribosomal protein L15, COG0197 ribosomal protein L16/L10E, COG0256 ribosomal protein L18, COG0090 ribosomal protein L2, COG0091 ribosomal protein L22, COG0089 ribosomal protein L23, COG0087 ribosomal protein L3, COG0088 ribosomal protein L4, COG0094 ribosomal protein L5, COG0097 ribosomal protein L6, COG0051 ribosomal protein S10, COG0100 ribosomal protein S11, COG0048 ribosomal protein S12, COG0099 ribosomal protein S13, COG0184 ribosomal protein S15P/S13E, COG0186 ribosomal protein S17, COG0185 ribosomal protein S19, COG0052 ribosomal protein S2, COG0092 ribosomal protein S3, COG0522 ribosomal protein S4 and related proteins, COG0098 ribosomal protein S5, COG0049 ribosomal protein S7, COG0096 ribosomal protein S8, COG0103 ribosomal protein S9, COG0172 seryl-tRNA synthetase, COG0441 threonyl-tRNA synthetase, COG0532 translation initiation factor 2 (GTPase), COG0180 tryptophanyl-tRNA synthetase, COG0525 valyl-tRNA synthetase, COG0202 DNA-directed RNA polymerase α-subunit/40-kDa subunit, COG0085 DNA-directed RNA polymerase β-subunit/140-kDa subunit, COG0250 transcription antiterminator, COG0258 5′-3′ exonuclease (including N-terminal domain of Pol I), COG0592 DNA polymerase III β-subunit, COG0468 RecA/RadA recombinase, COG0550 topoisomerase IA, COG0459 chaperonin GroEL (HSP60 family), COG0533 metal-dependent proteases with possible chaperone activity, COG0201 preprotein translocase subunit SecY, COG0541 signal recognition particle GTPase, COG0552 signal recognition particle GTPase, COG0112 glycine hydroxymethyltransferase, COG0125 thymidylate kinase, COG0237 dephospho-CoA kinase, COG0575 CDP-diglyceride synthetase, COG0012 predicted GTPase

conserved and non-conserved sites within the descendant sequence of each species were identified. Because these ancient sequences have diverged to a great extent, only slightly more than a third (~37%) of the sites within the ancestral sequence could be reconstructed. At sequence positions for which no ancestral residue could be assigned, it was assumed that residues within none of the descendant sequences were conserved. The frequency of each amino acid within conserved and non-conserved residues of the sequence set in each species could then be determined.

RESULTS

Conserved sequence elements for the 26 species were pooled to determine frequencies of each amino acid in those positions; the same was done for the non-conserved sequence elements. Six amino acids (glycine, histidine, leucine, proline, arginine, and tryptophan) were more frequent in conserved than non-conserved sequence elements; the remaining 14 amino acids were more frequent in non-conserved sequence elements (Table III).

The relative mutability of the 20 amino acids has been determined empirically by several investigators, starting with Dayhoff *et al.* (1). Jones *et al.* (11) later updated the mutability estimates of Dayhoff *et al.* (1) for a much larger set of protein sequences, whereas Gonnet *et al.* (12) based their mutability estimates on a set of sequences similar to that of Jones *et al.* (11) but using a modification of the Dayhoff approach. Depending on the data set and approach, some variations occur in the relative mutability ranking (Table IV). Nonetheless, seven amino acids (valine, glutamine, isoleucine, threonine, alanine, serine, and asparagine) consistently fall within the top half, and seven (tryptophan, cysteine, phenylalanine, tyrosine, glycine, proline, and arginine) fall within the bottom half of the ranking. There is, however, a lack of consensus as to whether the remaining amino acids (leucine, lysine, histidine, methionine, and glutamic and aspartic acids) are of high or low mutability. Accordingly, amino acids are assigned high, low, or undetermined mutability in Table III.

Based on both their relative mutabilities and their relative frequencies in conserved and non-conserved sequence elements, the following three amino acids may be inferred to have changed in frequency in the protein set since the LUA: cysteine, tyrosine, and phenylalanine (Table III). Because all three of these amino acids are of low mutability and are more abundant in non-conserved than conserved residues, they must have increased in frequency over time (Table I). Although valine, being of high mutability and occurring more frequently in conserved than non-conserved sequence elements, also satisfies the criteria summarized in Table I, its difference in frequency between these subsets is not statistically significant as determined using a chi-square test. The remaining amino acids either lack consensus regarding their relative mutability (see above) or fall into one of the two categories in Table I for which no inferences may be made; glycine, proline, arginine, and tryptophan are of low mutability and are more frequent in conserved than non-conserved residues, whereas alanine, isoleucine, glutamine, serine, and threonine are of high mutability and are less frequent in conserved than non-conserved residues.

The frequencies of cysteine, tyrosine, and phenylalanine within conserved residues are 0.0039, 0.0231, and 0.0331, respectively (Table III). Because of their low mutability, the frequencies of these amino acids within conserved residues provide an upper limit on their frequencies within this protein set in the LUA. By comparison, the frequencies of cysteine, tyrosine, and phenylalanine within the protein set as a whole are 0.0074, 0.0297, and 0.0374, respectively. It can therefore be inferred that the frequency of cysteine has doubled within this protein set between the LUA and today, whereas that of tyrosine has increased at least 29% and phenylalanine at least 13%.

Given these findings, we sought to determine whether the frequency of these three amino acids increased to an even greater extent within the modern whole-genome protein sets (*i.e.* proteomes) than within the ancient protein set. To this

TABLE III

Frequency of each amino acid in conserved and non-conserved sequence residues and in the entire protein set, pooled among the 26 species

Amino acids that consistently fall within the top half of the mutability ranking (see Table IV) are assigned high (H) relative mutability; those consistently in the bottom half, low (L) relative mutability; otherwise, undetermined (?) relative mutability.

Amino acid	Conserved	Non-conserved	Protein set	Conserved / Non-conserved	Relative mutability
Ala	0.0814	0.0821	0.0820	0.99	H
Cys	0.0039	0.0085	0.0074	0.45	L
Asp	0.0561	0.0551	0.0553	1.02	?
Glu	0.0779	0.0784	0.0782	0.99	?
Phe	0.0331	0.0388	0.0374	0.85	L
Gly	0.1320	0.0562	0.0738	2.35	L
His	0.0208	0.0192	0.0195	1.09	?
Ile	0.0593	0.0697	0.0673	0.85	H
Lys	0.0611	0.0799	0.0755	0.77	?
Leu	0.1079	0.0847	0.0901	1.27	?
Met	0.0128	0.0265	0.0233	0.48	?
Asn	0.0241	0.0402	0.0365	0.60	H
Pro	0.0629	0.0360	0.0423	1.74	L
Gln	0.0167	0.0375	0.0327	0.45	H
Arg	0.0710	0.0592	0.0620	1.20	L
Ser	0.0261	0.0563	0.0493	0.46	H
Thr	0.0385	0.0520	0.0488	0.74	H
Val	0.0801	0.0783	0.0787	1.02	H
Trp	0.0113	0.0097	0.0100	1.17	L
Tyr	0.0231	0.0317	0.0297	0.73	L

TABLE IV

Rank order of relative mutability of the amino acids (from most to least mutable) based on empirical data of Dayhoff et al. (1), Jones et al. (11), and Gonnet et al. (12)

	Dayhoff	Jones	Gonnet
Most mutable	Asn	Ser	Ser
	Ser	Thr	Ala
	Asp	Asn	Thr
	Glu	Ile	Gln
	Ala	Ala	Lys
	Thr	Val	Val
	Ile	Met	Glu
	Met	His	Asn
	Gln	Asp	Ile
	Val	Gln	Leu
	His	Arg	Met
	Arg	Glu	Asp
	Pro	Lys	Arg
	Lys	Pro	His
	Gly	Leu	Gly
	Tyr	Phe	Phe
	Phe	Gly	Pro
	Leu	Tyr	Tyr
	Cys	Cys	Cys
Least mutable	Trp	Trp	Trp

end, the mean frequency of each amino acid within the ancient protein set and within the proteomes was compared. Data on the proteomic frequency of these amino acids were taken from the Proteome Analysis Database (13). The mean frequency of cysteine within the ancient protein set is 0.0074 compared with 0.0099 in the proteomes, the frequency of tyrosine is 0.0297 *versus* 0.0335, and the frequency of phenylalanine is 0.0375 *versus* 0.0437. It is apparent, therefore, that the frequency of these three amino acids within modern proteomes has increased even more than within the set of ancient proteins itself.

To gain insight on whether cysteine, tyrosine, and phenylalanine might still be increasing in frequency today, we determined whether they are present in modern proteomes at frequencies predicted by neutral evolution. The neutral theory of molecular evolution predicts that an amino acid within a proteome should eventually reach an equilibrium frequency determined primarily by the number of codons assigned to that amino acid, adjusted for the nucleotide composition of its codons and the nucleotide composition of the genomic coding sequences (14). The probability of observing amino acid j in a specific genome is given by $p_j = \lambda(\sum_i x_i y_i z_i)$, where i represents each codon assigned to amino acid j; x_i, y_i, and z_i represent the frequency of occurrence of the first, second, and third nucleotides, respectively, of codon i within coding sequences of that genome; and λ is a constant such that the sum over all amino acids is equal to one. The normalization constant λ compensates for probabilities assigned to stop codons.

Using genomic coding sequence nucleotide frequency data derived from the Codon Usage Database (15), the frequencies of cysteine, tyrosine, and phenylalanine in the proteome of each species predicted by neutral evolution were determined (Table V). The observed frequency of cysteine is significantly less than that predicted in all 26 species ($p \ll 0.01$), the mean over all species being one-third of that predicted. In contrast, the observed frequencies of tyrosine is less than predicted in only 15 of the species ($p = 0.28$, which is not statistically significant), and the mean observed frequency of tyrosine,

TABLE V

Percent frequency of cysteine, tyrosine, and phenylalanine observed and predicted by neutral evolution in proteomes of 26 species

Species abbreviations are as follows: *Aquifex aeolicus*, Aae; *Archaeoglobus fulgidus*, Afu; *Aeropyrum pernix*, Ape; *Bacillus subtilis*, BAC; *Chlamydia pneumoniae*, CLA; *Campylobacter jejuni*, Cje; *Deinococcus radiodurans*, Dra; *Escherichia coli K12*, ENT; *Helicobacter pylori J9*, HPY; *Halobacterium* sp. *NRC-1*, Hbs; *Haemophilus influenzae*, Hin; *Mycoplasma pneumoniae*, MYC; *Methanococcus jannaschii*, Mja; *Methanobacterium thermoautotrophicum*, Mth; *Mycobacterium tuberculosis*, Mtu; *Neisseria meningitidis*, Nme; *Pseudomonas aeruginosa*, Pae; *Pyrococcus horikoshii*, Pyr; *Rickettsia prowazekii*, Rpr; *Borrelia burgdorferi*, SPI; *Saccharomyces cerevisiae*, Sce; *Synechocystis* PCC6803, Ssp; *Thermoplasma acidophilum*, Tac; *Thermotoga maritima*, Tma; *Vibrio cholerae*, Vch; *Xylella fastidiosa*, Xfa.

Species	Cysteine		Tyrosine		Phenylalanine	
	Proteome observed	Proteome predicted	Proteome observed	Proteome predicted	Proteome observed	Proteome predicted
Aae	0.0079	0.0266	0.0415	0.0357	0.0516	0.0260
Afu	0.0118	0.0309	0.0365	0.0296	0.0459	0.0251
Ape	0.0094	0.0311	0.0335	0.0222	0.0275	0.0209
BAC	0.0080	0.0301	0.0348	0.0376	0.0449	0.0321
CLA	0.0160	0.0328	0.0326	0.0457	0.0474	0.0463
Cje	0.0122	0.0296	0.0368	0.0593	0.0600	0.0535
Dra	0.0067	0.0267	0.0230	0.0136	0.0316	0.0127
ENT	0.0117	0.0332	0.0286	0.0298	0.0389	0.0293
HPY	0.0110	0.0303	0.0368	0.0449	0.0542	0.0394
Hbs	0.0075	0.0263	0.0255	0.0148	0.0312	0.0128
Hin	0.0104	0.0321	0.0315	0.0479	0.0447	0.0456
MYC	0.0075	0.0292	0.0323	0.0440	0.0559	0.0385
Mja	0.0128	0.0277	0.0438	0.0516	0.0426	0.0401
Mth	0.0121	0.0283	0.0322	0.0285	0.0365	0.0225
Mtu	0.0088	0.0295	0.0208	0.0148	0.0296	0.0152
Nme	0.0103	0.0288	0.0298	0.0275	0.0412	0.0237
Pae	0.0100	0.0271	0.0254	0.0144	0.0356	0.0136
Pyr	0.0063	0.0303	0.0384	0.0387	0.0460	0.0322
Rpr	0.0110	0.0285	0.0389	0.0603	0.0488	0.0536
SPI	0.0073	0.0270	0.0429	0.0608	0.0619	0.0511
Sce	0.0130	0.0285	0.0380	0.0453	0.0450	0.0385
Ssp	0.0100	0.0335	0.0291	0.0343	0.0401	0.0347
Tac	0.0060	0.0294	0.0464	0.0330	0.0470	0.0272
Tma	0.0071	0.0286	0.0358	0.0332	0.0519	0.0262
Vch	0.0105	0.0334	0.0296	0.0352	0.0407	0.0347
Xfa	0.0119	0.0340	0.0262	0.0277	0.0347	0.0291
Mean	0.0099	0.0298	0.0335	0.0358	0.0437	0.0317

0.0335, is close to that predicted, 0.0358. For phenylalanine, the observed frequency is higher than predicted in 25 species ($p \ll 0.01$), the mean observed frequency, 0.0437, being ~40% higher than predicted. Therefore, the observed frequency of cysteine is less than, and of phenylalanine is greater than, that predicted by neutral evolution, whereas that of tyrosine agrees with the prediction of neutral evolution.

DISCUSSION

It is generally assumed that those amino acids believed to have been absent from the prebiotic environment were added to the genetic code later, as enzymes for their biosynthesis evolved (16). Thus, very early versions of the code would have included only prebiotically-available amino acids. Because cysteine, tyrosine, and phenylalanine are absent from simulations of the prebiotic environment of the Earth (17), they are commonly held to be late additions to the genetic code. Although we do not propose a specific mechanism for addition of these amino acids to the evolving primitive code, we do make the assumption that codon reassignments would have occurred in a fashion that introduced them into proteins gradually, because the impact upon protein structure of introducing these amino acids en masse was more likely to be detrimental than beneficial (18). Specifically, these amino acids most likely adopted codons that occurred infrequently within coding sequences. This idea is consistent with the fact that both cysteine and tyrosine share four-codon blocks with at least one stop codon; it is quite possible that the code had only recently evolved to use those codons to specify other amino acids (through modification of existing tRNAs) when cysteine and tyrosine "captured" them.

Consequently, we propose that upon their introduction into the code, these three amino acids would have gone from being non-existent to being rare within early coded proteins. Furthermore, because of the distinct physicochemical properties of these amino acids, the majority of subsequent coding sequence mutations introducing them into proteins presumably would have been deleterious, causing their increase in frequency to be gradual (that of cysteine especially so). Because our data indicate that these three amino acids increased in frequency between the LUA and today, they must not have reached their equilibrium frequencies by the time of the LUA. According to this scenario, the under-representation

of these amino acids in the LUA relative to today is consistent with their late addition to the genetic code.

It has conventionally been assumed that the time between the origin of proteins and today has been sufficient for all amino acids to reach their equilibrium frequencies and therefore, that an observed frequency of an amino acid distinct from that predicted by neutral evolution is evidence of some strict requirement of protein structure or function that places unusual selection on that amino acid (14). However, because our findings suggest that at the time of the LUA, cysteine, tyrosine, and phenylalanine had yet to reach equilibrium frequencies, change of amino acid composition toward that predicted by neutral evolution may be a process requiring very long time periods. Indeed, the observation that the frequency of cysteine is so much lower than that predicted by neutral evolution in modern proteomes may be evidence that the increase in usage of this particular amino acid has been especially gradual over evolution. Consequently, the possibility that even today cysteine continues to move toward its equilibrium frequency through neutral evolution, as the vast range of all possible sequence space is gradually searched, cannot be ruled out. On the other hand, over time phenylalanine has become more frequent in proteins than predicted by neutral evolution. In fact, it is possible that the frequency of phenylalanine, too, will increase further with evolution. In any case, positive selection for phenylalanine has caused any initial rarity of this amino acid in the earliest proteins to be overcome. The same may be argued for tyrosine, the observed frequency of which does not differ significantly from that predicted by neutral evolution.

Although our approach did not produce evidence for a change in frequency of any of the other 17 amino acids over the course of evolution, this does not imply that no other amino acids have changed in frequency. Using our rationale, it is not possible to reach a definite conclusion regarding the change in frequency (or lack thereof) of those amino acids of high mutability that are less frequent in conserved than non-conserved positions and those of low mutability that are more frequent in conserved than non-conserved positions. Moreover, our ability to make inferences was limited by the lack of consensus on the relative mutability of six amino acids (see Table III and Table IV). It is therefore possible that amino acids other than cysteine, tyrosine, and phenylalanine have increased in frequency since the LUA. With the increase in frequency of these three (and perhaps other) amino acids, there must have been a concomitant decrease in frequency of at least one other amino acid. Because valine is of low mutability and is present at greater frequency in conserved than non-conserved sequence elements (although not to a statistically significant extent), it may indeed have decreased in frequency over time. An alternative approach will be required to determine with certainty which amino acids other than cysteine, tyrosine, and phenylalanine have in fact changed in frequency over evolution.

It is not immediately evident how amino acid composition and structure have co-evolved in the ancient protein set investigated. Studies of protein evolution suggest that structure and function can be well conserved even as protein sequence diverges extensively (see Ref. 19, but see Ref. 20 for a contrary view). However, evolution of amino acid composition may have impacted structure in newly arising proteins of the proteome. Each amino acid has a specific predisposition to occur in different secondary structures, *i.e.* in α-helices, β-sheets, or random coils (21, 22), and negative selection preserving structure would have been relatively relaxed in this later protein set. Further investigation will be required to elucidate structural consequences of changes in proteomic amino acid composition.

* The computational facility utilized for this work was obtained with funds provided by the Department of Defense through MEDCOM at Fort Detrick, MD (to J. R. F.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Supported in part by predoctoral traineeships from National Institutes of Health Grant 2T32GM07388-22 and from National Science Foundation Grant DGE 9972930.

§ To whom correspondence should be addressed. Tel.: 609-258-3927; Fax: 609-258-2759; E-mail: jrfresco@princeton.edu.

REFERENCES

1. Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978) *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed) Vol. 5, Suppl. 3, pp. 345–352, National Biomedical Research Foundation, Washington, D. C.
2. Woese, C. R. (1998) The universal ancestor. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 6854–6859
3. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29,** 22–28
4. Kyrpides, N. C., Overbeek, R., and Ouzounis, C. A. (1999) Universal protein families and the functional content of the last universal common ancestor. *J. Mol. Evol.* **49,** 413–423
5. Olsen, G. J., Woese, C. R., and Overbeek, R. (1994) The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* **176,** 1–6
6. Fitz-Gibbon, S. T., and House, C. H. (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.* **27,** 4218–4222
7. Rivera, M. C., Jain, R., Moore, J. E., and Lake, J. A. (1998) Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 6239–6344
8. 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
9. Eck, R. V., and Dayhoff, M. O. (1966) *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed) pp. 166–169, National Biomedical Research Foundation, Silver Spring, MD
10. Felsenstein, J. (1989) PHYLIP (Phylogeny Inference Package, Version 3.2, *Cladistics* **5,** 164–166
11. Jones, D. R., Taylor, W. R., and Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8,** 275–282
12. Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992) Exhaustive matching of the entire protein sequence database. *Science* **256,** 1443–1445

13. Apweiler, R., Biswas, M., Fleischmann, W., Kanapin, A., Karavidopoulou, Y., Kersey, P., Kriventseva, E. V., Mittard, V., Mulder, N., Phan, I., and Zdobnov, E. (2001) Proteome analysis database: online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. *Nucleic Acids Res.* **29,** 44–48
14. King, J. L., and Jukes, T. H. (1969) Non-Darwinian evolution. *Science* **164,** 788–798
15. Nakamura, Y., Gojobori, T., and Ikemura, T. (2000) Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.***28,** 292
16. Wong, J. T. (1975) A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **72,** 1909–1912
17. Miller, S. L. (1987) Which organic compounds could have occurred on the prebiotic earth? *Cold Spring Harbor Symp. Quant. Biol.* **52,** 17–27
18. Osawa, S., Jukes, T. H., Watanabe, K., and Muto, A. (1992) Recent evidence for evolution of the genetic code. *Microbiol Rev.* **56,** 229–264
19. Chothia, C., and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5,** 823–826
20. Wood, T. C., and Pearson, W. R. (1999) Evolution of protein sequences and structures. *J. Mol. Biol.* **291,** 977–995
21. Chou, P. Y., and Fasman, G. D. (1974) Conformational parameter for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry* **13,** 211–217
22. King, R. D., and Sternberg, M. J. (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* **5,** 2298–2310