Do Older Taxa Have Older Proteins?

Ricardo Ferreira[a,b,*], Frederico J. S. Pontes[a], Benício de Barros Neto[a], and Patrícia M. A. Farias[b]

[a] Departamento de Química Fundamental/CCEN, Universidade Federal de Pernambuco, 50670-901, Recife, Pernambuco, Brasil. Fax: (55)-81-3271-8442.
E-mail: rferreira100@yahoo.com

[b] Departamento de Biofísica e Radiobiologia/CCB, Universidade Federal de Pernambuco, 50670-901, Recife, Pernambuco, Brasil

* Author for correspondence and reprint requests

Z. Naturforsch. **59 c**, 454–458 (2004); received September 25/November 13, 2003

We have confirmed through an enlarged set of 728 species with 10,000 or more compiled codons, and a subset of 237 species with at least 50,000 compiled codons, that the mean values of a previously described index Φ [the mean value of the ratio between the relative (G, C) content of Class II and Class I codons, where G and C are guanine and cytosine] decrease monotonically across five large taxa, viz archaea, bacteria, eukaryotes (excluding metazoa), metazoa (excluding vertebrates) and vertebrates. It is proposed that these main taxa diverge successively from an ancestral progenome along lines which have persisted over long periods of time, leading to a primordial non-symmetrical phylogenetic tree. Further divergence, *i.e.* from eukaryotes to plants, fungi and protozoans, has followed symmetrical branching with approximately equal numbers of replacements and fixations. A statistical analysis of the Φ values of twelve distinct proteins, distributed over more than one thousand species belonging to the five main groups, was made to verify whether older taxa have older proteins. This supposition was confirmed for the first four taxa, but it was inconclusive for the last pair, metazoa/vertebrates.

Key words: Proteins, Compiled Codons, Evolution

Introduction

According to the Evolutionary Theory any present day species, such as, for example, the fig-tree (*Ficus carica*), is not older than, say, a modern *Homo sapiens*, since both are separated by the same time span from a common ancestor, and during the divergence process the two have undergone approximately the same number of mutations and fixations. In the same way, the amino acid sequence for species which have diverged from a common ancestor have experienced an approximately equal number of replacements, most of which corresponding to neutral mutations (Kimura, 1968). The number of fixations, *i.e.* surviving replacements, should also be approximately constant. This should lead to *symmetrical* phylogenetic trees, such that the "same" protein as it occurs in distinct modern species differs by approximately the same number of residues from the corresponding ancestor proteins (Doolittle, 1979).

If we restrict ourselves to this evolutionary pattern, the very notion that a given species may be older, or newer, than another species must be qualified. Thus, present day *E. coli* has branched-out from *Salmonella* only 100 million years ago, and it is not older than, say, ray-finned fishes (*Actinopterygeans*).

In this paper we argue that early taxa with a common progenome have followed evolutionary lines which have persisted over long periods of time, leading to five large groups of organisms, namely archaeabacteria, bacteria, eukaryotes (excluding metazoa), metazoa (excluding vertebrates) and vertebrates. These old lines must have accommodated many polymorphic neutral alleles, with few opportunities for fixations. This implies that at this stage the phylogenetic trees of these large taxa were quite *asymmetrical*.

Our conclusion stems from an analysis of the existence of two distinct classes of aminoacyl t-RNA synthetases, the enzymes that bind the amino acids to their cognate t-RNAs (Eigen and Winkler-Oswaltisch, 1981).

0939–5075/2004/0500–0454 $ 06.00

Results

The two classes of aminoacyl t-RNA synthetases

It is now well established that the aminoacyl t-RNA synthetases are descendent from two distinct ancestral enzymes (Eriani *et al.*, 1990, 1995; Cusack *et al.*, 1990; Nagel and Doolittle, 1991), even though the transfer RNAs probably have a common origin. Although there are claims in the literature that the two classes of synthetases may have originated simultaneously (Rodin and Ohno, 1995; Pouplona and Schimmel, 2001); there are previous evidences that Class II synthetases were incorporated earlier than Class I enzymes into the translation apparatus (Hartman, 1995; Ferreira and Cavalcanti, 1997). The latter model has been strenghtened by more recent results (Cavalcanti *et al.*, 2000) showing that the mean chemical distances resulting from mutations that do not change the class of the involved amino acids are much smaller than those of mutations which do change the class of the amino acids, so that the former mutations should be responsible for most of the recognized minimization of the genetic code.

These findings strongly support models for the origin and evolution of the genetic code according to which new amino acids were incorporated to an original version of the code containing fewer than twenty amino acids, following duplication and divergence of previously existing synthetases and t-RNAs. This is, in fact, the process proposed by Crick in his "frozen accident" hypothesis (Crick, 1965).

Enlargement of the amino acid cast of proteins has now been experimentally shown to be feasible by the generation of a bacterium with a 21 amino acid genetic code (Mehl *et al.*, 2003). The newly incorporated amino acid is *p*-aminophenylalanine (pAF), and its corresponding synthetase belongs to Class I.

The index Φ: its uses and limitations as a tracer in molecular evolution

We have found (Cavalcanti and Ferreira, 2001) that the mean value of the ratio between the relative (G, C) content of Class II and Class I codons, defined by the index:

$$\Phi = \frac{\dfrac{(C + G)_{II}}{(C + G + A + U)_{II}}}{\dfrac{(C + G)_{I}}{(C + G + A + U)_{I}}}$$

where G, C, A and U are the known bases of DNA, decreases monotonically, in a statistically significant way, within a 95% confidence interval in the sequence archaeabacteria, bacteria, eukaryotes (excluding metazoa), metazoa (excluding vertebrates), and vertebrates.

In a previous communication (Ferreira and Cavalcanti, 1997) using a series of CUTG database releases (Nakamura *et al.*, 2000) we included 530 species with 10,000 or more compiled codons, and, as a subset, 152 species with 50,000 or more compiled codons. We have now extended our calculations to cover 728 and 237 species, respectively, from the same database (Nakamura *et al.*, 2000).

Thus, from the protein coding regions of the genomes of 728 species with at least 10,000 compiled codons [15 archaea, 379 bacteria, 192 eukaryotes (excluding metazoa), 69 metazoa (excluding vertebrates) and 73 vertebrates], we have found the mean Φ values given in Table I. For the 237 species with 50,000 or more compiled codons [13 archaea, 138 bacteria, 47 eukaryotes (excluding metazoa), 11 metazoa (excluding vertebrates) and 28 vertebrates] the calculated mean Φ values are in Table I too.

These results are not significantly different, for either set, from the previously reported values

Table I. Mean Φ values calculated for species with at least 10,000 and 50,000 compiled codons.

	10,000 Codons compiled	50,000 Codons compiled
Taxa	Φ (mean ± standard error)	Φ (mean ± standard error)
Archaea	1.331 ± 0.017	1.337 ± 0.019
Bacteria	1.264 ± 0.004	1.267 ± 0.005
Eukaryotes	1.209 ± 0.005	1.198 ± 0.005
Metazoa	1.169 ± 0.007	1.167 ± 0.008
Vertebrates	1.110 ± 0.005	1.108 ± 0.005

(Cavalcanti and Ferreira, 2001), which shows that the index is statistically robust. So far as the five taxa are concerned, this is an indication that the observed behavior reflects a general trend in the early course of evolution.

This behavior is adequately accounted for if, at the stage in which the translation apparatus was being developed, there existed a large excess of the relative (G, C) content of Class II codons over Class I codons. Thus, the earliest form of life had higher values of the Φ index, and these higher values are still shown by prokaryotes.

We have found that the values of Φ for eukaryotes cannot be further splitted into protozoans, plants and fungi. We interpret this limitation as indicative that, starting with an eukaryotic common ancestor, further divergence leading to the new three taxa followed a symmetrical phylogenetic tree with an approximately equal number of replacements and fixations. As a result, the differences in exon composition between protozoa and fungi, protozoa and plants, and plants and fungi are approximately constant.

The hypothesis of a predominance in early times of the relative (G, C) content of codons of Class II amino acids over those of Class I is not completely free from difficulties. From the point of view of Organic Chemistry purines, especially adenines, are easier to synthesize than pyrimidines (Zubay and Mice, 2001).

There are, on the other hand, experimental results showing that poly(C)-directed oligomerization of guanine is much easier than poly(U)-directed oligomerization of adenine (Joyce, 1987; Orgel, 2002).

The early predominance of (G, C)-rich ribotides over (A, U)-rich ones is also predicted by models of the kinetics of oligoribotide growth (Ferreira, 1987; Ferreira and Coutinho, 1993).

The question whether the last common ancestor of bacteria is less ancient than archaeabacteria has not been definitely answered. Cavalier-Smith (2001), for example, has proposed that archaeabacteria are not older than 850 million years, whereas Schopf (2002) believes that Doolittle's dating of 2,000 million years for the origin of the eukaryotes is correct (Doolittle *et al.*, 1996), which implies that archaeabacteria are much older than indicated by Cavalier-Smith's estimation.

Our figures show that archaeabacteria are older than bacteria, which is the conclusion put forward by Carl Woese in his classical study (Woese, 1987). The fact that some archaeabacteria genomes contain few introns does not contradict this view. The controversy between the "introns late" (Cavalier-Smith, 1991) and "introns early" (Doolittle, 1978; Gilbert, 1987) has been solved by a compromise: about 30–40% of present day intron positions were originally present in the progenomes (ancestral eukaryotes), while almost all the remaining intron positions correspond to introns added to the progenomes (Gilbert *et al.*, 1998). The classification of archaeabacteria as the oldest taxa so far described gains support from the painstaking work of Doolittle and coworkers (Doolittle *et al.*, 1997). We quote from their paper: "The majority of the archaeabacterial sequences are not compatible with currently accepted views of the Tree of Life, which cluster the archaeabacteria with the eukaryotes. Instead they are either outliers or mixed with the eubacterial orthologs"; they go on to say that "these two groups (archaea and eubacteria) may have diverged between 3 and 4 billion years ago".

We can also dismiss the possibility that the higher mean value of Φ for archaeabacteria is due to their introduction having taken place in high temperature environment, which should make them richer in amino acids with high thermophylic rank (Di Giulio, 2000). However, of the 10 amino acids with the highest thermophylic rank, six, including the most thermophylic of all, arginine, belong to Class I and contribute instead to decrease the mean value of Φ in that class.

An analysis of individual proteins

We have tested whether the behavior shown by the mean Φ values of exons is repeated for individual proteins. We computed the Φ values of the "same" protein in various species belonging to five large taxa. Twelve proteins were used in the calculations: enolase, aldolase, lactate dehydrogenase, pyruvate kinase, aldehyde dehydrogenase, glyceraldehyde 3-phosphate, aconitase, malate dehydrogenase, succinate dehydrogenase, malate sintase, isocitrate liase and fumarase. The first six are enzymes which act in glycolysis, while the other six are involved in Krebs cycle. The set is distributed among 107 archaea, 572 bacteria, 362 eukaryotes, 162 metazoa and 80 vertebrates, representing 1,383 codon defining segments in all. The distributions of the Φ values for the five taxa are visual-

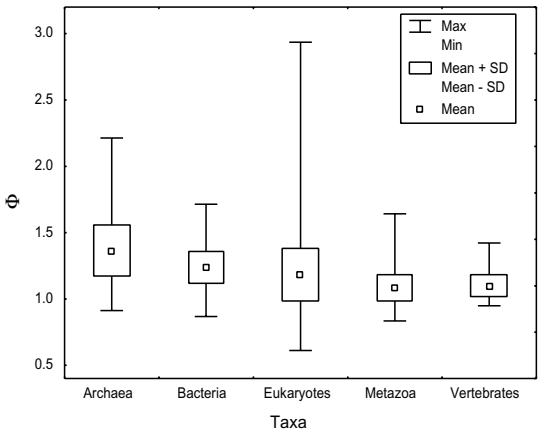

Fig. 1. Distribution of calculated Φ values among the five taxas.

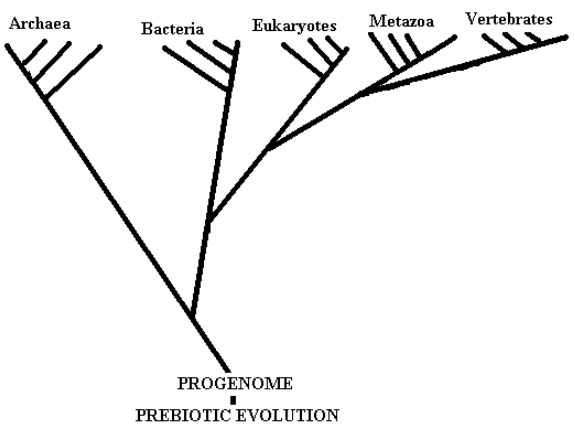

Fig. 2. Schematic phylogenetic tree of life for evolution of the taxa.

ized in the box-and-whisker plot of Fig. 1. The corresponding mean values are given in Table II, along with the results of *t*-tests performed on the four pairs of adjacent taxa. The mean Φ values for the first four taxa are in the expected decreasing order, with extremely significant *p*-values for the respective pairwise *t*-tests. For the metazoa-vertebrates comparison, on the other hand, the test is inconclusive. The mean Φ values for these two taxa are in fact statistically indistinguishable. Perhaps this might be ascribed to the relatively small number of degrees of freedom (240) on which this particular comparison is based.

Discussion and Conclusion

It is quite clear that the gradual change in codon composition gauged by the steady decrease of the values of Φ along the main taxa cannot be due to entirely randomic processes but express part of the dynamics of natural selection.

Thus, if all mutations and fixations responsible for this gradual change were strictly random, in the sense that they occur with the same frequency independently of the codon position in the exons, it comes directly from probability theory that, starting with a species population of any codonic distribution, after a sufficiently large number of events the population will attain the most probable distribution, which is the codonic distribution of the genetic code itself. Now, the genetic code contains 32 Class II codons with 54 (G, C) bases, and 29 Class I codons with 40 (G, C) bases. The value of Φ for the code is therefore

$$\Phi = \frac{(54)/(96)}{(40)/(87)} = 1.223$$

Our figures show that this value was reached, in the course of evolution, already with the earliest eukaryotes and has continued to decrease for metazoa and vertebrates. This is incompatible with pure genetic drift and call for some process of natural selection.

Table II. Comparison of mean Φ values calculated for the twelve enzymes in 1,383 codon defining segments (CDS).

Taxa	Mean	Standard deviation	Number of CDS's	Degrees of freedom	*p*-Value of *t*-test[a]
Archaea	1.363	0.195	107	–	–
Bacteria	1.236	0.121	572	677	0.0000
Eucaryotes	1.183	0.201	362	932	0.0000
Metazoa	1.084	0.102	162	522	0.0000
Vertebrates	1.100	0.083	80	240	0.2291

[a] The *p*-values in this column refer to *t*-tests performed between the taxon shown in the line and that in the previous line.

On the basis of this study and its main conclusions we wish to propose a schematic tree of life, congruent with it. This tree is shown in Fig. 2.

Acknowledgement

We wish to acknowledge the Brazilian Agency CNPQ for financial support, and Dr. André Ricardo Cavalcanti (Princeton University) for helpful suggestions.

Cavalcanti A. R. O. and Ferreira R. (2001), On the relative content of G, C bases in codons of amino acids corresponding to class I and II aminoacyl t-RNA synthetases. Orig. Life Evol. Biosph. **31**, 257–269.

Cavalcanti A. R. O., Ferreira R., and Neto B. B. (2000), On the classes of aminoacyl-tRNA synthetases and the error minimization in the genetic code. J. Theor. Biol. **204**, 15–20.

Cavalier-Smith T. (1991), Intron phylogeny: a new hypothesis. Trends in Genetics **7**, 145–148.

Cavalier-Smith T. (2001), Obcells as proto-organisms. J. Mol. Evol. **53**, 555–585.

Crick F. H. C. (1965), The origin of the genetic code. J. Mol. Evol. **38**, 367–378.

Cusack S., Berthet-Colominas C., Hartlein M., Nassar N., and Leberman R. (1990), A second class structure revealed by X-ray analysis of *Escherichia coli* seryl-t-RNA synthetase. Nature **347**, 249–255.

Di Giulio M. (2000), The late stage of genetic code structuring took place at a high temperature. Gene **261**, 188–195.

Doolittle R. F. (1978), Genes in pieces – were they ever together. Nature **272**, 581–582.

Doolittle R. F. (1979), Protein evolution. In: The Proteins IV (Neurath H. and Hill R., eds.). Academic Press, New York, pp. 1–118.

Doolittle R. F., Cho G., Feng F., and Tsang S. (1996), Determining divergence times of the major kingdoms of living organisms with a protein clock, Science **271**, 470–476.

Doolittle R. F., Cho G., and Feng F. (1997), Determining divergence times with a protein clock: update and reevaluation. Proc. Natl. Acad. Sci. USA **94**, 13028–13033.

Eigen M. and Winkler-Oswaltisch R. (1981), Transfer RNA. The early adaptor. Naturwissenschaften **68**, 217–228.

Eriani G., Delarue M., Poch O., Gangloff J., and Moras D. (1990), Partition of t-RNA synthetases into two classes based on mutually exclusive sets of sequence motif. Nature **347**, 203–206.

Eriani G., Cavaleri J., Martin F., Ador L., Reees B., Thierry J. C., Gangloff J., and Moras D. (1995), The class II aminoacyl t-RNA synthetases and their active sites. J. Mol. Evol. **40**, 499–508.

Ferreira R. (1987), A two-substrate Michaelis-Menten model for the growth of self-replication polymers. J. Theor. Biol. **128**, 289–295.

Ferreira R. and Coutinho K. R. (1993), Simulation studies of the self-replicating oligoribotides, with a proposal to a peptide-assisted stage. J. Theor. Biol. **164**, 291–305.

Ferreira R. and Cavalcanti A. R. O. (1997), Vestiges of early molecular processes leading to the genetic code. Orig. Life Evol. Biosph. **27**, 397–403.

Gilbert W. (1987), The exon theory of genes. Cold Spring Harbor Symp. Quant. Biol. **52**, 901–905.

Gilbert W., Souza S. J., Long M., Klein R. J., and Roy S. (1998), Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins. Proc. Natl. Acad. Sci. USA **95**, 5094–5099.

Hartman H. (1995), Speculations on the origin of the genetic code. J. Mol. Evol. **40**, 541–544.

Joyce G. F. (1987), Non-enzymatic template directed synthesis of informational macromolecules. Cold Spring Harbor Symp. Quant. Biol. **52**, 41–51.

Kimura M. (1968), Evolutionary rate at the molecular level. Nature **217**, 624–626.

Mehl J., Ryan A., Christopher A., Santoro S. W., Wang L., Martin A. B., King D. S., Horn D. H., and Schultz P. G. (2003), Generation of a bacterium with a 21 amino acid genetic code. J. Am. Chem. Soc. **125**, 935–939.

Nagel G. M. and Doolittle R. F. (1991), Evolution and relatedness in two aminoacyl t-RNA synthetases families. Proc. Natl. Acad. Sci. USA **88**, 8121–8125.

Nakamura Y., Gojobori T., and Ikemura T. (2000), Codon usage tabulated from international DNA Sequence Databases: status for the year 2000. Nucl. Acids. Res. **28**, 292.

Orgel L. E. (2002), private communication.

Pouplona L. R. and Schimmel P. (2001), Two classes of t-RNA synthetases suggested by sterically compatible dockings on tRNA acceptor sterm. Cell **104**, 191–193.

Rodin S. N. and Ohno S. (1995), Two types of aminoacyl t-RNA synthetases could be originally encoded by complementary strands of the same nucleic acid. Orig. Life Evol. Biosph. **25**, 565–589.

Schopf J. W. (1993), Microfossils of the early archean aprox. chert: New evidence of the antiquity of life. Science **260**, 640–646.

Schopf J. W. (2002), private communication.

Wetzel R. (1995), Evolution of the aminoacyl t-RNA synthetase and the origin of the genetic code. J. Mol. Evol. **40**, 545–550.

Woese C. R. (1987), Bacterial evolution. Microbiol. Rev. **51**, 221–227.

Zubay G. and Mice T. (2001), Prebiotic synthesis of nucleotides. Orig. Life Evol. Biosph. **31**, 87–102.