Gene Ther Mol Biol Vol 4, 313-322. December 1999.

Glycine clock: Eubacteria first, Archaea next, Protoctista, Fungi, Planta and Animalia at last

Research Article

Edward N. Trifonov

Department of Structural Biology, The Weizmann Institute of Science, Rehovot 76100, Israel

Correspondence: E. N. Trifonov, Department of Structural Biology, The Weizmann Institute of Science, Rehovot 76100, Israel. **Fax**: +972 8 934 2653; E-mail edward.trifonov@weizmann.ac.il

Key words: evolutionary trees, triplet code, earliest proteins, amino-acid chronology, amino-acid composition, molecular evolution, codon chronology, primordial soup, multiple alignments,

Received: 15 April 1999; accepted 25 April 1999

Summary

Twenty-five different single-factor criteria and hypotheses about chronological order of appearance of amino acids in the early evolution are summarized in consensus ranking. All available knowledge and thoughts about origin and evolution of the genetic code are thus combined in a single list where the amino acids are ranked in descending order, starting with the earliest ones:

G, A, D, V, P, S, E, L, T, I, N, F, H, K, R, Q, C, M, Y, W

One may expect that in the composition of the ancient proteins the earliest amino acids would dominate. Indeed, when homologous prokaryotic and eukaryotic protein sequences are aligned, the most frequent residue amongst matching amino acids (presumably, what remains of the common ancestor sequence) is glycine that makes about 14% vs. glycine content of 6-7% in modern proteins. The glycine content of the matching residues may, then, serve as a measure of the time (glycine clock) since the separation of compared species. This approach is applied to 370 pairwise alignments of protein sequences from over 100 species of 6 major kingdoms. The evolutionary tree is derived, where the kingdoms separate consecutively from the central stem in the order: Eubacteria (13.5% G at the moment of separation), Archaea (11.5%), Protoctista (10.5%), Fungi (9%), Planta/Animalia (8%), largely consistent with common knowledge on the evolution of the kingdoms. The glycine content, thus, may serve as a time label that allows the tracing back of the separation of any two species with potential accuracy of the order of 50 to 100 million years, all the way to the very origin of species.

I. Introduction

The molecular clocks of which many sophisticated versions had been developed since original suggestion by Zuckerkandl and Pauling (1962), suffer from numerous drawbacks (see, e. g., Doolittle, 1997; Ayala et al., 1998), especially when applied to very early molecular events. In particular, the evolutionary rates are not constant, the distance estimates are influenced by horizontal transfer, and double (multiple) replacements are difficult to account for. The quantitative evaluations of similarity in the sequence comparisons become unreliable when too little of a common ancestor is left in the sequences. Moreover, the sequence dissimilarity indicates evolutionary distance between the sequences, but the time direction remains uncertain, resulting in so-called unrooted evolutionary trees. It would be highly desirable to find some internal property(ies) of the sequences that would indicate their evolutionary age. One such property is suggested by the recently derived chronological ranking of amino acids, order of their appearance on the early evolutionary scene (Trifonov and Bettecken, 1997; Trifonov, 1999). The earliest amino acids should have been overrepresented in the earliest proteins, in which case mere amino-acid composition could serve as the indicator of the age of the protein. This approach, however, can not be used in as straightforward way, since all extant proteins are of the same age, if one assumes that the proteins originate from their immediate and distant ancestors, rather than formed *de novo* (Zuckerkandl, 1976). One way to evaluate the amino-acid composition of the proteins of the distant past is to compare (align) related sequences from evolutionary distant species and take the composition of shared residues. As it is described below, the "common" composition of eukaryotic and prokaryotic sequences (evolving separately about 3 Gyrs), indeed, is strongly biased towards the earliest amino acids, in particular, glycine. This suggests to use the glycine content as measure of time (glycine clock) passed since separation of the species, to construct the rooted evolutionary tree.

II. Results and discussion

A. Amino-acid composition of early proteins

The earliest form of the triplet code has been recently reconstructed, consisting of 10 codons and 7 respective amino acids: ala, asp, gly, pro, ser, thr and val (Trifonov and Bettecken, 1997). The reconstruction was based on natural expandability of $(GCT)_n$ sequences, and on universal $(GCU)_n$ pattern hidden in mRNA sequences (Lagunez-Otero and Trifonov, 1992). This suggested that the very first triplets were GCU and it’s 9 point change derivatives. The reconstruction of the above list of the earliest amino acids was based on the experiments of S. L. Miller (1987), on chemical simplicity of the amino acids and on association with more ancient class II aminoacyl-tRNA synthetases. Inspection of the table of the triplet code revealed a striking correspondence between all these residues and the GCU-derived codons (Trifonov and Bettecken, 1997). This gives reason to believe that the earliest proteins, perhaps, long time before the separation of eukaryotes from prokaryotes, had been built from the above 7 ancient residues. At later stages, with appearance of other amino acids the domination of the seven, surely, was compromised. However, one could expect that even at the stage of separation eukaryotes-prokaryotes some of the ancient residues still prevailed. Further insight into the amino-acid chronology is provided by adding to the analysis four more criteria of the amino-acids' evolutionary age, in addition to the above three: frequency of occurrence of various amino acids in modern proteins, stability of the codon-anticodon interactions, chemical inertness of amino acids, and the GCU triplet-based list of the amino acids, as an independent criterion. Ranking analysis of the seven "chronologies" suggested by these criteria (Trifonov, 1999) resulted in the following list of the amino acids, in descending order of their appearance on the evolutionary scene: ala, gly, ser, pro, val, thr, leu, asp, ile, glu, asn, phe, lys, arg, gln, cys, his, met, trp and tyr. The earliest proteins, therefore, would be expected to contain less of the latest amino acids, say, gln, cys, his, met, trp and tyr. As a matter of fact, these residues, indeed, are least frequent even in extant proteins (see **Figure 1**), but the early proteins, perhaps, had even less of these residues. This is checked by alignment of prokaryotic and eukaryotic sequences and comparing amino-acid composition of the common parts (points) to the composition of modern eukaryotic and prokaryotic proteins. In extension of an earlier work (Trifonov, 1998) this analysis is performed on 70 arbitrarily chosen functionally different aligned sequence pairs (**Table 1**), scoring total 5551 matching residues. The actual scores and amino-acid compositions in % are presented in the **Table 2** and in the **Figure 1** (under "common"). In this Figure the composition values for prokaryotic and eukaryotic proteins (two upper plots) are taken from Arques and Michel, (1996). The histograms presented in the **Figure 1** show, first of all, that in the common (about 3 billion years old) eukaryotic-prokaryotic material the gly residues are significantly more frequent (about twice) than in modern proteins. This major bias is observed even when only 10 sequence pairs are taken for the analysis. In the **Table 2** amino-acid compositions for 7 different sets, 10 sequence pairs each, are presented (sequence Nos. 1-10, 11-20, 21-30, ... 61-70 of the **Table 1**). In all cases the domination of glycine is obvious: 12.7 to 15.5 % versus 6 - 7% in modern proteins. To be sure that the bias is not due to overrepresentation of some species, *E. coli* in particular (33 sequence pairs), two sets have been assembled, one dominated by *E. coli* sequences (set 7) and another one - with *E. coli* sequences underrepresented (set 6). The content of gly is found to be high in both cases. Total of 27 different prokaryotic species and 32 eukaryotic species are represented in the 70 sequence pairs analyzed (**Table 1**). The effect, therefore, is general, apparently reflecting, indeed, the amino-acid composition of the proteins at the moment of separation between prokaryotes and eukaryotes.

If the ratios of the occurrences in "common" to the occurrences in prokaryotes and in eukaryotes are considered, then two more amino acids appear on the top: asp and pro (about 20% excess). All three including glycine belong to the earliest alphabet. That is, the earliest amino acids have been still overrepresented at the time of separation eukaryotes-prokaryotes. Glycine, aspartic acid and proline are known to be the most specific residues for the turns of folded polypeptide chains (Kwasigroch *et al.*, 1996). Their unusual conservation, thus, indicates that the turns are no less important in maintaining conserved protein structure than alpha-helices and beta-sheets.

Another conspicuous feature of the "common" distribution (**Figure 1**) is an abrupt drop of composition values for the amino acids tyr, asn, his, gln, met, trp and cys. Five of them belong to the latest in the amino-acid chronology (Trifonov, 1999, and manuscript in preparation).

It appears, thus, that about 3 billion years back these "young" residues have been just entering the scene being, therefore, substantially less numerous than the "older" residues. Their share in the total, according to our data, was 10.7%, versus 30% for even distribution of amino acids. No such step in the amino-acid composition is observed in case of modern proteins (**Figure 1**, upper plots) though the "young" residues are underrepresented here as well. It appears, thus, that since the time of separation eukaryotes-prokaryotes the proportion of the "young" residues increased, apparently, in the process of their gradual accommodation and optimization of the protein composition. The proportion of the latest residues as well as excess of the earliest glycine residues may, thus, potentially serve for timing of the evolutionary bifurcations.

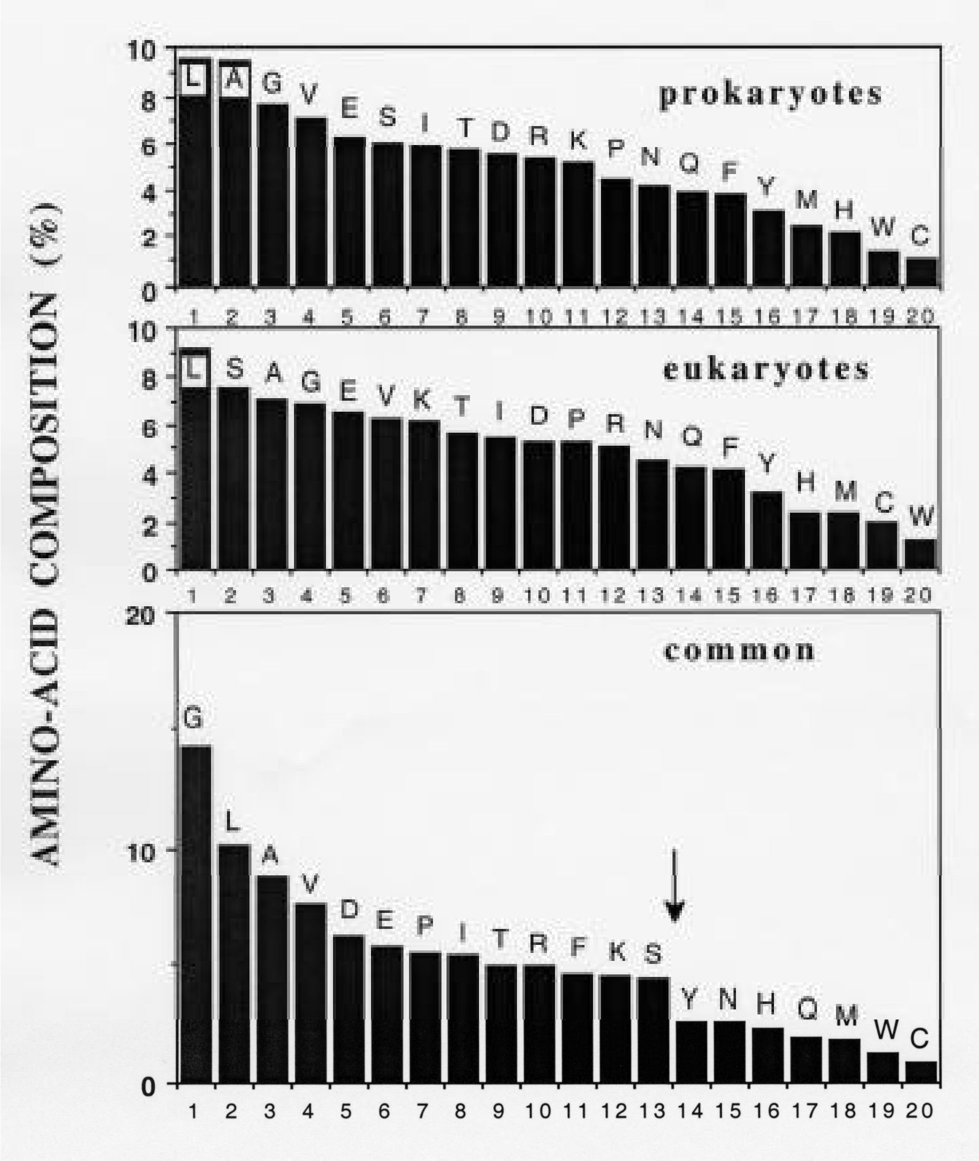


Figure 1. Amino-acid composition of matching residues in alignments of related prokaryotic and eukaryotic protein sequences ("common") as compared to modern proteins of prokaryotes and eukaryotes.

Exceptional status of glycine in molecular evolution has been indicated earlier in the study on the correlation of the evolutionary rate with the amino-acid composition (Graur, 1985). An "almost uninterchangeable" glycine was found to be "one of the most conserved amino acids". This also suggests higher content of glycine in the older, conserved proteins. Being the smallest amino acid glycine serves very much as a hinge in the polypeptide chain providing it with high flexibility. The conformational versatility would be of high importance in the early stages of protein evolution. Later on, perhaps, with advance in sophistication of the protein structure rather stability of the evolved conformations became important, and the glycine content eventually came down to the modest present level.

B. The amino-acid and codon chronology

More extended analysis involving 25 different amino-acid age criteria (manuscript in preparation) arrives to the chronology very similar to the one listed above. A vertical column on the left of the **Figure 2** represents the order of the amino acids, in which they, presumably, appeared on the evolutionary scene. All available knowledge and thoughts about origin and evolution of the genetic code are combined in this single list where the amino acids are ranked in descending order, starting with the earliest ones. The ranking is inevitably of rather poor accuracy. The typical differences in the calculated ranks as compared with the earlier 7-criteria list are 1-2 ranks.

Table 1. Aligned prokaryotic-eukaryotic protein sequence pairs.

Species	**Protein (gene)**	**Reference**
1. Escherichia coli human	thymidilate synthase --“--	Gene 150, 221, 1994
2. Halobact. cutirubrum C. elegans	hypothetical G-protein --“--	Gene 151, 153, 1994
3. Bacteroides fragilis maize	pyruvate dikinase --“--	Gene 151, 173, 1994
4. Flavobact. meningosepticum pig	prolyl endopeptidase	Gene 152, 103, 1995
5. Escherichia coli rabbit	phosphofructokinase ATP-dep phosphofructokinase	Gene 152, 181, 1995
6. Bacillus circulans Brugia malayi (nematode)	chitinase A3 chitinase	Gene 153, 147, 1995
7. Enterococcus faecium carrot	dihydrofolate reductase --“--	Gene 154, 7, 1995
8. Agrobact. tumefaciens X. laevis	Arginase --“--	Gene 154, 115, 1995
9. Escherichia coli human	ribosomal protein S1 --“-- , repeat 2	Gene 155, 231, 1995
10. Escherichia coli human	glutathione reductase --“--	Gene 156, 123, 1995
11. Escherichia coli mouse	ribose 5-phosphate isomerase --“--	Gene 156, 191, 1995
12. Escherichia coli tomato	RNase I RNase LE	Gene 158, 203, 1995
13. Clostridium acetobutylicum C. elegans	3-hydroxyacyl CoA dehydrogenase --“-- (F54C8.6)	Gene 160, 309, 1995
14. Alcaligenes Arabidopsis thaliana	Nitrilase --“--	Gene 161, 15, 1995
15. Escherichia coli Arabidopsis thaliana	adenine phosphorybosyltransferase --“--	Gene 161, 81, 1995
16. Pseudomonas Aspergillus nidulans	NAD-dep. formate dehydrogenase --“--	Gene 162, 99, 1995
17. Escherichia coli rat	arginyl-tRNA synthetase --“--	Gene 164, 347, 1995
18. Escherichia coli mouse	RNA polymerase subunit α RNA polymerase I/III AC40	Gene 167, 203, 1995
19. Escherichia coli C. elegans	RNA polymerase subunit α RNA polymerase III AC16	Gene 172, 211, 1996
20. B. stearothermophilus Plasmodium knowlesi	valine-tRNA synthetase --“--	Gene 173, 137, 1996
21. B. cereus rabbit	thermolysin microsomal endopeptidase	Gene 174, 135, 1996
22. B. subtilis mouse	inosine monophosphate dehydrogenase --“--	Gene 174, 209, 1996
23. B. subtilis human	methylenomycin A resistance protein glucose transporter type I	Gene 175, 223, 1996
24. Escherichia coli Aspergilus nidulans	NARK nitrate transporter CRNA nitrate transporter	Gene 175, 223, 1996
25. Lactobacillus sake Chinese hamster	SapT (sakacin synthesis) multidrug resistance protein	Gene 176, 55, 1996
26. Rhodobacter capsulatus Triticum aestivum	S-adenosylhomocysteine hydrolase --“--	Gene 177, 17, 1996
27. B. subtilis rat	3-methyladenine DNA glycosylase --“--	Gene 177, 229, 1996

28. Escherichia coli rice	Mrp (ATPase) EST D25016 (ATPase)	Gene 178, 97, 1996
29. Escherichia coli red alga	3-ketoacyl-acyl carrier prot. synthase --“--	Gene 182, 45, 1996
30. B. subtilis Arabidopsis thaliana	protoporphyrinogen oxidase --“--	Gene 182, 169, 1996
31. Pseudomonas putida human	glyoxalase I --“--	Gene 186, 103, 1997
32. Escherichia coli mouse	spermidine synthase --“--	Gene 187, 35, 1997
33. Escherichia coli rabbit	glutaredoxin --“--	Gene 188, 23, 1997
34. Zymomonas mobilis human	glyceraldehyde-3-phosphate DH --“--	Gene 188, 221, 1997
35. Zymomonas mobilis human	phosphoglycerate kinase --“--	Gene 188, 221, 1997
36. B. megaterium human	triosephosphate isomerase --“--	Gene 188, 221, 1997
37. Escherichia coli Brassica napus	phosphoenolpyruvate carboxykinase --“--	Gene 192, 235, 1997
38. B. subtilis rat	peptidylprolyl cis-trans isomerase --“--	Gene 193, 65, 1997
39. B. subtilis X. laevis	Arginase --“--	Gene 193, 157, 1997
40. Escherichia coli dog	signal peptidase I --“--	Gene 194, 249, 1997
41. Rhizobium leguminosarum D. discoideum	orotate phosphorybosyltransferase --“--	Gene 195, 329, 1997
42. B. subtilis human	myo-inositol 2-dehydrogenase biliverdin reductase	Gene 196, 209, 1997
43. Escherichia coli Schistosoma mansoni	cold-shock protein CSPA Y-box binding protein	Gene 198, 5, 1997
44. Streptococcus mutans tobacco	non-phosphorylating GAPN --“--	Gene 198, 237, 1997
45. Staphylococcus xylosus human	histone deacetylase (acuC) --“-- (HDm)	Gene 198, 275, 1997
46. Escherichia coli human	heat-shock protein HSP 60 --“--	Gene 199, 83, 1997
47. Escherichia coli human	porphobilinogen deaminase --“--	Gene 199, 231, 1997
48. Synechococcus barley	HemL protein --“--	Gene 199, 231, 1997
49. Escherichia coli D. melanogaster	RNA helicase --“--	Gene 199, 241, 1997
50. P. aeruginosa T. bruce i	mercuric reductase trypanothione reductase	Gene 200, 163, 1997
51. B. subtilis Geodia cydonium	alcohol dehydrogenase AidB-like protein	J. Mol. Evol. 47, 343, 1998
52. Legionella pneumophila bovine	Cu, Zn superoxide dismutase --“--	J. Mol. Biol. 274, 408, 1997
53. Thermus aquaticus mouse	DNA polymerase (5'-3' exonucl.domain) flap endonuclease (FEN-1)	J. Biol. Chem. 272, 28531, 1997
54. M. genitalium tobacco	uracil phpsphoribosyltransferase --“--	EMBO J. 17, 3219, 1998
55. Synechococcus elongatus Chlamydomonas reinhardtii	photosystem II RC domain --“--	J. Mol. Biol. 280, 1998

56. Rhodobacter capsulatus tobacco	uroporphyrinogen decarboxylase --“--	EMBO 17, 2463, 1998
57. Streptomyces hydrogenans Drosophila lebanonensis	3α,20β-hydroxysteroid dehydrogenase alcohol dehydrogenase	J. Mol. Biol. 282, 383, 1998
58. Escherichia coli C. elegans	transition metal transporter --“--	J. Biol. Chem. 272, 28485, 1997
59. T. thermophilus human	histidyl-tRNA synthetase --“--	J. Mol. Biol. 280, 847, 1998
60. Escherichia coli D. melanogaster	pspE HSP67Bb	J. Mol. Biol. 282, 195, 1998
61. Escherichia coli human	GTP-binding protein (FtsY) --“-- (SRα)	Gene 201, 37, 1997
62. Escherichia coli rabbit	trehalase --“--	Gene 202, 69, 1997
63. Escherichia coli D. melanogaster	parvulin Dodo protein	Gene 203, 89, 1997
64. Escherichia coli rat	aminopeptidase N --“--	Biochemistry 37, 686, 1998
65. Escherichia coli Brugia malayi	asparaginyl-tRNA synthetase --“--	EMBO J. 17, 2947, 1998
66. Escherichia coli human	glutathione S-transferase --“--	J. Mol. Biol. 271, 135, 1998
67. Escherichia coli rice	thioredoxin glutaredoxin	J. Mol. Biol. 281, 949, 1998
68. Escherichia coli human	glutaredoxin thioredoxin	J. Mol. Biol. 281, 949, 1998
69. Escherichia coli Flaveria trinervia	phosphoenolpyruvate carboxylase --“--	J. Mol. Evol. 46, 107, 1998
70. Escherichia coli human	periplasmic cyclophilin cyclophilin A1	EMBO J. 17, 2463, 1998

Despite this uncertainty, due to consensus nature of the chronology it has several important properties not visible in individual rankings by any of the initial criteria. The conclusion of the earlier GCU-based theory on the structure of the earliest code is confirmed: all 7 earliest amino acids are, indeed, found at the top of the consensus chronology (G, A, D, V, P, S and T). Ten amino acids of the Miller's imitation of primordial soup are all ranked as topmost (G, A, D, V, P, S, E, L, T, I). This result is especially important, since it confirms that, indeed, the experimental conditions chosen by Miller are close to the primordial ones, and that the first amino acids acquired by the emerging life were synthesized abiotically.

The consensus order of appearance of the 20 amino acids on the evolutionary scene also reveals a unique and simple chronological organization of 64 codons, that could not be figured out from individual criteria: new codons appear in complementary pairs, with the complement recruited from the codon repertoire of the earlier or simultaneously appearing amino acids. The resulting codon chronology also reveals that of alternative codon-anticodon pairs the most stable ones appear first, if not all together.

Contrary to the GCU-based theory of the origin of the code, it is glycine rather than alanine that appears at the top of the list. Actually, they appear simultaneously, within the accuracy of the ranking (manuscript in preparation). The apparent contradiction, however, rather suggests a correction to the GCU-model. As it was indicated in the paper on the GCU theory (Trifonov and Bettecken, 1997), the GCC triplet and its point change derivatives correspond to the same seven earliest amino acids. The first codons, thus, could be, indeed, GCC and GGC, for alanine and glycine, respectively, in accordance with the chronology displayed in the **Figure 2**. This pair of codons has been suggested as the earliest ones 20 years ago by Eigen and Schuster (1978). What is important for the elaboration in the next section - the glycine is one of the earliest amino acids. It apparently took over at some time in the early evolution becoming a dominant residue (see **Figure 1**).

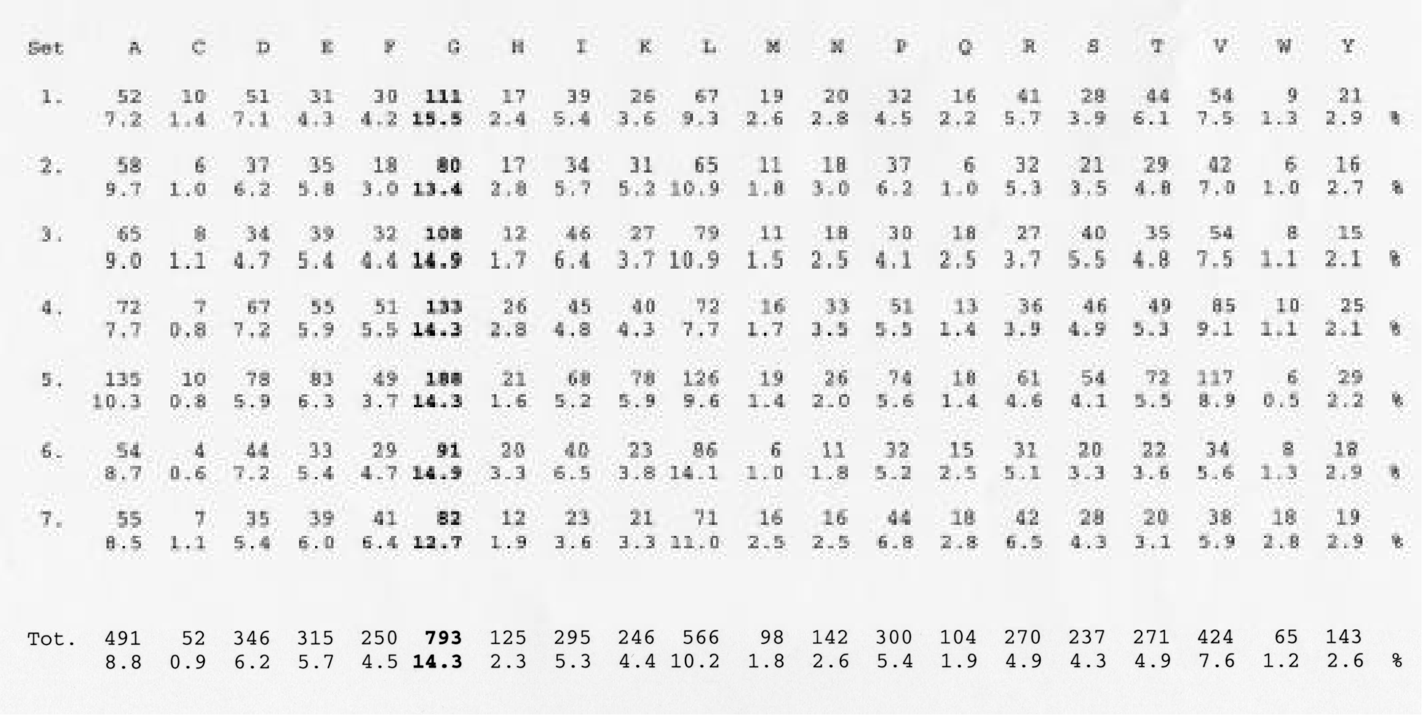
C. Glycine clock and evolutionary tree for six major kingdoms.

The calculations similar to those made for the prokaryotes and eukaryotes, as presented in the **Tables 1** and **2**, are performed for sequence pairs from 6 major kingdoms: eukaryotes (Protoctista, Fungi, Planta and Animalia) and prokaryotes (Eubacteria and Archaea). Total 370 sequence pairs are analyzed, and the average contents of the glycine amongst the shared residues are calculated for each of 15 groups of the kingdom-to-kingdom sequence comparisons. The functionally diverse sequences are taken from literature, basically, on the random basis. They represent as large variety of species, as exemplified by the **Table 1**. In the **Table 3** the derived values are presented, together with actual scores (in brackets, glycine/total). The number of sequence pairs used for the analysis is indicated as well (italics). The errors are calculated on the assumption that the scatter in the actual scores of glycines follows normal distribution with STD equal to square root of the score.

The highest contents of glycine among the shared residues of the aligned sequences is observed for Eubacteria (see **Table 3**). The respective % GLY values vary between 12.1 ± 1.2% and 14.8 ± 0.6% with the average 13.7 ± 0.3%. If only eukaryotes are taken for the alignments with the eubacterial protein sequences, as in the **Table 2**, the average % GLY value from the new set of the sequences is 1460/10602 = 13.8 ± 0.4%, to compare with 14.3 ± 0.5% for the earlier set (**Table 2**), indistinguishable within the error bars. The % GLY values for Archaea, compared to four eukaryotic kingdoms, vary between 11.3 ± 1.0% and 13.3 ± 1.5%, with the average 11.7 ± 0.6%, clearly lower than the above average value for Eubacteria. That would correspond to a later separation of the Archaea from eukaryotes, some time after Eubacteria. The % GLY value for separation Archaea-Eubacteria, on the other hand, is close to the separation level for Eubacteria, as it would be expected, 12.8 ± 0.9% vs. 13.7 ± 0.3%. Similarly, the % GLY values for later separations of Protoctista, Fungi and Planta are progressively lower, while comparisons of their sequences with older kingdoms give higher % GLY values, corresponding, respectively, to the separation times of the latter.

The % GLY values are arranged in the **Table 3** in such a way that the line averages of the values provide the branching level of % GLY for respective kingdoms. Of 15 kingdom-to-kingdom % GLY values only 3 (< 32% of 15) are more than 1 STD off the respective averages, which, thus, justifies the assumed normal distribution of the % GLY estimates. The evolutionary tree based on the % GLY values presented in the **Table 3** is shown on the **Figure 3**. This tree is very much consistent with the trees derived from molecular clock calculations (Feng *et al.*, 1997; Doolittle, 1997; Otsuka *et al.*, 1999). If the time separation between branchings of plants and of Eubacteria is taken equal 2 Gyrs, 1% GLY corresponds to about 350 Myrs. This provides an approximate calibration of the glycine clock. At this early stage of the development of the glycine clock the linear calibration is an understandable simplification. Both the **Table 3** and the **Figure 3** represent the first estimates of the branchings of the major kingdoms, based on only 370 sequence pairs. The number of the sequences can be substantially increased (say, to many thousands), so that the tree would be subject of further improvements towards better accuracy. However, as the current error bars indicate, the overall topology of the basic tree will most likely stay unchanged.

Table 2. Amino-acid composition of common residues in eukaryotic-prokaryotic sequence alignments



Set	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
1.	52	10	51	31	30	**111**	17	39	26	67	19	20	32	16	41	28	44	54	9	21	
	7.2	1.4	7.1	4.3	4.2	**15.5**	2.4	5.4	3.6	9.3	2.6	2.8	4.5	2.2	5.7	3.9	6.1	7.5	1.3	2.9	%
2.	58	6	37	35	18	**80**	17	34	31	65	11	18	37	6	32	21	29	42	6	16	
	9.7	1.0	6.2	5.8	3.0	**13.4**	2.8	5.7	5.2	10.9	1.8	3.0	6.2	1.0	5.3	3.5	4.8	7.0	1.0	2.7	%
3.	65	8	34	39	32	**108**	12	46	27	79	11	18	30	18	27	40	35	54	8	15	
	9.0	1.1	4.7	5.4	4.4	**14.9**	1.7	6.4	3.7	10.9	1.5	2.5	4.1	2.5	3.7	5.5	4.8	7.5	1.1	2.1	%
4.	72	7	67	55	51	**133**	26	45	40	72	16	33	51	13	36	46	49	85	10	25	
	7.7	0.8	7.2	5.9	5.5	**14.3**	2.8	4.8	4.3	7.7	1.7	3.5	5.5	1.4	3.9	4.9	5.3	9.1	1.1	2.1	%
5.	135	10	78	83	49	**188**	21	68	78	126	19	26	74	18	61	54	72	117	6	29	
	10.3	0.8	5.9	6.3	3.7	**14.3**	1.6	5.2	5.9	9.6	1.4	2.0	5.6	1.4	4.6	4.1	5.5	8.9	0.5	2.2	%
6.	54	4	44	33	29	**91**	20	40	23	86	6	11	32	15	31	20	22	34	8	18	
	8.7	0.6	7.2	5.4	4.7	**14.9**	3.3	6.5	3.8	14.1	1.0	1.8	5.2	2.5	5.1	3.3	3.6	5.6	1.3	2.9	%
7.	55	7	35	39	41	**82**	12	23	21	71	16	16	44	18	42	28	20	38	18	19	
	8.5	1.1	5.4	6.0	6.4	**12.7**	1.9	3.6	3.3	11.0	2.5	2.5	6.8	2.8	6.5	4.3	3.1	5.9	2.8	2.9	%
Tot.	491	52	346	315	250	**793**	125	295	246	566	98	142	300	104	270	237	271	424	65	143	
	8.8	0.9	6.2	5.7	4.5	**14.3**	2.3	5.3	4.4	10.2	1.8	2.6	5.4	1.9	4.9	4.3	4.9	7.6	1.2	2.6	%

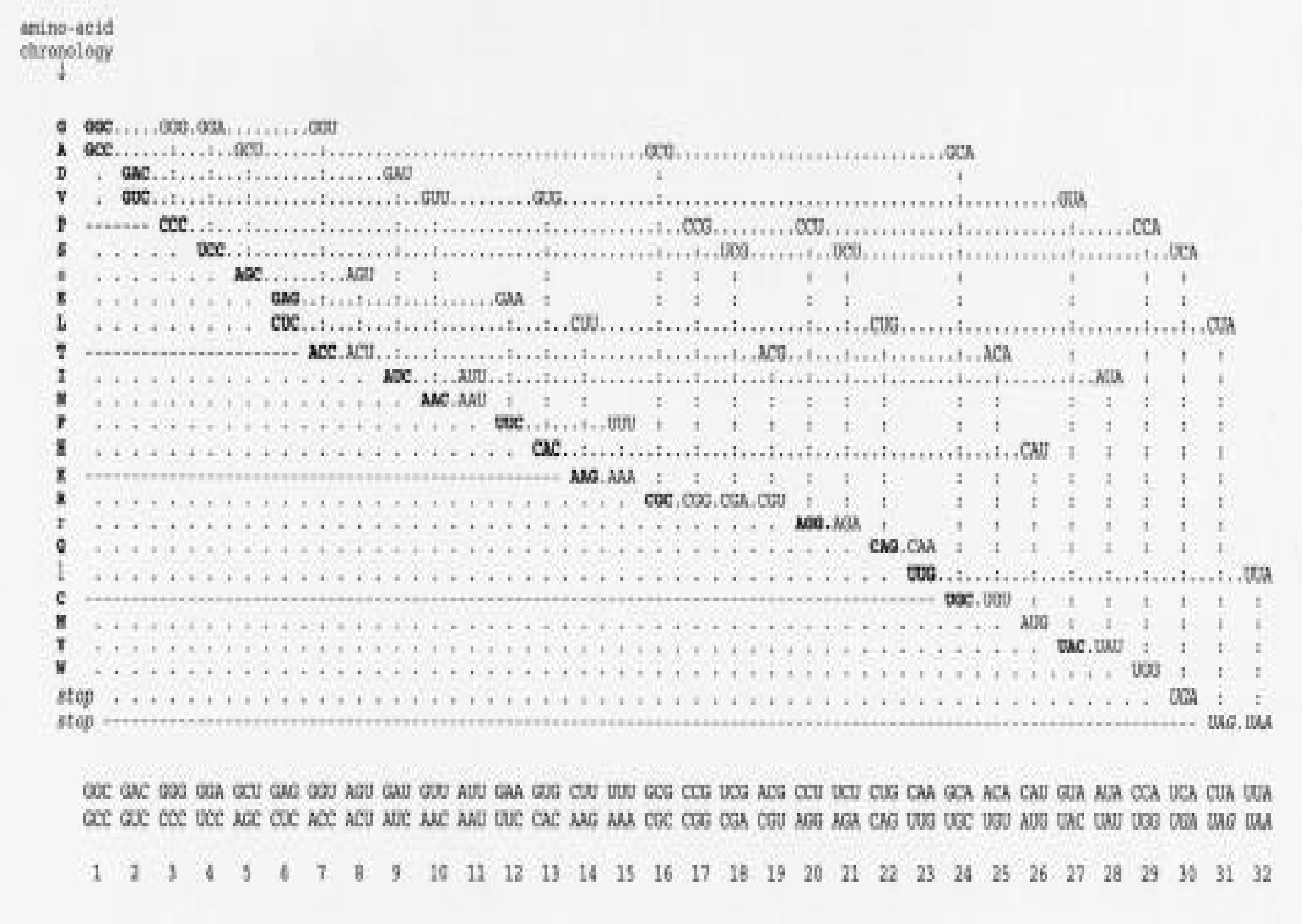

amino-acid chronology
G
A
D
V
P
S
s
E
L
T
I
N
F
H
K
R
r
Q
l
C
M
Y
W
stop
stop
GUC GAC GGG GGA GCU GAG GGU AGU GAU GUU AUU GAA GUG CUU UUU GCG CCG UCG ACG CCU UCU CUG CAA GCA ACA CAU GUA AUA CCA UCA CUA UUA
GCC GUC CCC UCC AGC CUC ACC ACU AUC AAC AAU UUC CAC AAG AAA CGC CGG CGA CGU AGG AGA CAG UUG UGC UGU AUG UAC UAU UGG UGA UAG UAA
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32

Figure 2. Chronology of 32 codon pairs. The amino-acid chronology is calculated as average ranking based on 25 different criteria. The codon chronology is one simple way of arranging the 64 triplets in accordance with the amino-acid chronology. Of alternative codons those which make most stable codon-anticodon pairs are engaged first (bold). In this case there is always a complementary triplet available, of the codon repertoires for earlier amino acids.

Table 3. Contents of shared glycine (%) in kingdom-to-kingdom protein sequence alignments

	ANIMALIA	PLANTA	FUNGI	PROTOCTISTA	ARCHEA	Branching level
PLANTA	8.1± 0.6 (193/2194, 25)					8.1± 0.6 (193/2194, 25)
FUNGI	8.88±0.4 (573/6479, 70).	9.1±0.7 (179/1977, 23)				8.9±0.3 (752/8456, 93)
PROTOCTISTA	11.1±1.1 (98/879, 11)	9.8±0.8 (156/1595, 10)	11.4±1.0 (137/1200, 11)			10.6±0.5 (391/3674, 32)
ARCHEA	11.3±1.0 (128/1133, 18)	11.7±1.7 (49/418, 12)	11.3±1.0 (132/1170, 19)	13.3±1.5 (82/616, 8)		11.7 ±0.6 (391/3337, 57)
EUBACTERIA	14.8±0.6 (584/3935, 63)	13.1±0.7 (313/2381, 21)	13.4±0.6 (468/3502, 46)	12.1±1.2 (95/784, 10)	12.8±0.9 (187/1462, 23)	13.7±0.3 (1647/12064, 163)

It is noteworthy that the glycine clock approach (or, presumably, any other approach based on the content of the earliest amino acids) apparently provides both evolutionary distance (in % GLY time units in this case) and directionality (the larger the branching % GLY value the older the separation event). This would allow to construct a detailed rooted tree, with further subdivisions of the kingdoms and potential resolution of 50 to 100 Myrs, the higher the more sequences are taken for the alignments. The technique is especially promising in dating the earliest separations where sensitivity of the classical molecular clock is low. The tree in the **Figure 3** is presented in its simplest form, with the central stem from which the respective kingdoms separate in the chronological order as indicated. Animalia rather than Planta are chosen to crown the tree, to reflect the obvious trend displayed by the tree - from the simplest to the most complex. Indeed, anuclear prokaryotes separate first, followed by the nucleated eukaryotes. The eukaryotes, on the other hand, progress from unicellular to multicellular, differentiated organisms. In a way, at each stage the simpler forms separated from the stem that continued to evolve to yet more complex forms. In that sense the common ancestor of all kingdoms though, perhaps, as simple as Eubacteria at the moment of their separation, was omnipotent having carried all elements that later evolved into the higher complexity of younger kingdoms. The higher evolutionary potential stayed associated with the main stem at every next branching. The branches of the kingdoms in the **Figure 3** are not continued to the top of the tree, to the typical and common modern 6-7% of GLY, although this is implied, in order to better reflect the linear succession of the branching events.

Apart from appealing simplicity of the glycine clock, its directionality and applicability to the earliest branchings, this technique is substantially less dependent on the effects of horizontal transfer and variations in the evolutionary rates. These are averaged over large number of sequences that are taken for the calculations.

III. Sequences and methods

The aligned prokaryotic-eukaryotic sequence pairs are collected from literature, irrespective of the alignment technique chosen by the authors of the original papers. To ensure random choice of the sequences, all alignments published in **Gene**, volumes 150 to 200, have been taken for the ensemble in the **Table 1**, total 50 sequence pairs. Additional 20 pairs are collected from various sources, on random basis as well. For all 440 sequence comparisons used in this work only those sequence pairs are taken which are part of multiple alignments of no less than 4 sequences in each. Matching residues are scored which are separated by no more than 4 non-matching residues, with no gaps (local sequence similarity ≥ 33%). Wherever possible, the sequence pairs are taken to represent as broad variety of species as the sequence data allow.

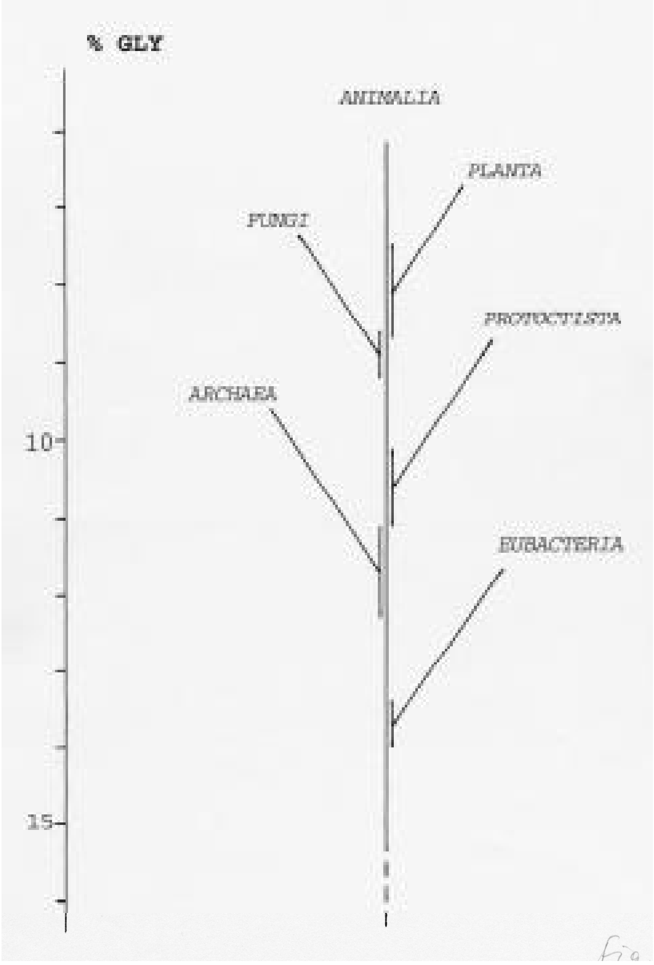


Figure 3. Evolutionary tree of major kingdoms, according to glycine clock estimates. The glycine content % GLY corresponds to the proteins existing at the moment of separation of respective kingdoms. The vertical bars at the separation points indicate current uncertainty of the estimates, dependent on the amount of the sequences compared.

References

Arques, D. G., and Michel, C. J. (**1996**) A complementary circular code in the protein coding genes. **J. Theor. Biol.** 182, 45-58.

Ayala, F. J., Rzhetsky, A., and Ayala F. J. (**1998**) Origin of the metazoan phyla: molecular clocks confirm paleontological estimates. **Proc. Natl. Acad. Sci. USA** 95, 606-611.

Doolittle, W. F. (**1998**) Fun with genealogy. **Proc. Natl. Acad. Sci. USA** 94, 12751-12753.

Eigen, M., and Schuster, P. (1978) The hypercycle. A principle of natural self- organization. Part C: The realistic hypercycle. **Naturwissenschaften** 65, 341-369.

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

Graur, D. (**1985**) Amino acid composition and the evolutionary rates of protein-coding genes. **J. Mol. Evol**. 22, 53-62.

Kwasigroch, J.-M., Chomilier, J., and Mornon, J.-P. (**1996**) A global taxonomy of loops in globular proteins. **J. Molec. Biol.** 259, 855-872.

Lagunez-Otero, J., and Trifonov, E. N. (**1992**) mRNA periodical infrastructure complementary to the proof-reading site in the ribosome. **J. Biomol. Struct. Dynam.** 10, 455-464.

Miller, S. L. (1987) Which organic compounds could have occurred on the prebiotic earth. **Cold Spr. Harb. Symp. Quant. Biol.** 52, 17-27.

Otsuka, J., Terai, G., and Nakano, T. (**1999**) Phylogeny of organisms investigated by the base-pair changes in the stem regions of small and large ribosomal subunit RNAs. **J. Mol. Evol.** 48, 218-235.

Trifonov, E. N. (**1998**) How basics of protein evolution could help the gene finding. Proceedings of the First International Conference on Bioinformatics of Genome Regulation and Structure BGRS'98, Novosibirsk - Altai Mountains, August 24-31, 1998, ICG, Novosibirsk, v.2. **Bioinformatics of Genome Structure**, pp. 266-268

Trifonov, E. N. (**1999**) Elucidating sequence codes: three codes for evolution. **Annals NY Acad. Sci.,** in press

Trifonov, E. N., and Bettecken, T. (**1997**) Sequence fossils, triplet expansion, and reconstruction of earliest codons. **Gene** 205, 1-6.

Zuckerkandl, E. (**1975**) The appearance of new structures and functions in proteins during evolution. **J. Mol. Evol.** 7, 1-57.

Zuckerkandl, E., and Pauling, L. (**1962**) Molecular disease, evolution and genetic heterogeneity. In: Kasha, M., and Pullman, B., (eds.) **Horizons in Biochemistry**. Academic Press, New York, pp. 189-225.

Ed Trifonov