Orig Life Evol Biosph (2006) 36:559–565
DOI 10.1007/s11084-006-9042-5

Primordia Vita. Deconvolution from Modern Sequences.

Edward N. Trifonov · Idan Gabdank · Danny Barash · Yehoshua Sobolevsky

Published online: 22 November 2006

Abstract Evolution of the triplet code is reconstructed on the basis of consensus temporal order of appearance of amino acids. Several important predictions are confirmed by computational sequence analyses. The earliest amino acids, alanine and glycine, have been encoded by GCC and GGC codons, as today. They were succeeded, respectively, by *A*- and *G*-series of amino acids, encoded by pyrimidine-central and purine-central codons. The length of the earliest proteins is estimated to be 6–7 residues. The earliest mRNAs were short G+C-rich molecules. These short sequences could have formed hairpins. This is confirmed by analysis of modern prokaryotic mRNA sequences. Predominant size of detected ancient hairpins also corresponds to 6–7 amino acids, as above. Vestiges of last common ancestor can be found in extant proteins in form of entirely conserved short sequences of size six to nine residues present in all or almost all sequenced prokaryotic proteomes (omnipresent motifs). The functions of the topmost conserved octamers are not involved in the basic elementary syntheses. This suggests an initial abiotic supply of amino acids, bases and sugars.

Keywords abiotic synthesis · ancient binary alphabet of proteins · earliest mini-genes · earliest mRNA · earliest proteins · evolution of triplet code · last common ancestor · origin of life · reconstruction of ancient sequences

The origin and early evolution of the triplet code are, presumably rather late in the chain of events initiated by the origin of Life. However, if anything certain could be said about the first steps of the triplet code this would likely be of great value for extrapolating back to the

Presented at: *National Workshop on Astrobiology: Search for Life in the Solar System*, Capri, Italy, 26 to 28 October, 2005.

E. N. Trifonov (✉) · D. Barash · Y. Sobolevsky
Genome Diversity Center, Institute of Evolution, University of Haifa, Haifa 31905, Israel
e-mail: trifonov@research.haifa.ac.il

I. Gabdank · D. Barash
Department of Computer Science, Ben Gurion University of the Negev,
P.O. Box 653, Be'er Sheva 84105, Israel

Life origin. Recent reconstruction of the evolution of the triplet code on the basis of consensus chronology of amino acids (Trifonov 2000, 2004) opened a broad perspective for elucidation of further important details of the early evolution. Several non-trivial predictions are formulated on the basis of the reconstructed evolutionary chart of codons (Trifonov et al. 2001; Trifonov 2006), and confirmed by analysis of modern sequence data.

In this work several observations along these lines are brought together in an attempt to outline the spectrum of the earliest proteins and their functions. The analysis strongly points to a generous abiotic supply of elementary building units of Life, amino acids and nucleobases, in the beginning of Life.

Results and Discussion

The first principles of the evolution of the triplet code (Trifonov 2004), suggested by the consensus evolutionary temporal order of amino acids are: (1) Abiotic start, (2) Primacy of thermostability, (3) Complementarity of codons and of early mRNA, (4) Processivity of codon acquirements, each having its precursor, and (5) Codon capture at latest stages. The codon chart, its latest update, is shown in the Figure 1. It is built on the basis of total 82 different criteria for temporal order of amino acids in evolution, of which 60 have been used for construction of previous versions of the chart (Trifonov 2000, 2004), and 22 new criteria are taken from additional independent sources (references can be provided by request). As it was found in previous work (Trifonov 2000, 2004), the final chart is not sensitive to filtering procedures, so that even simple averaging of the chronology vectors results in the temporal order identical to the filtered data, within error bars of the estimates. Thus, for the latest update of the chart (Figure 1) a simple summation of the vectors is used. The only filtering applied is removal of earlier chronologies of the same author. The consensus temporal order of amino acids is shown on the top of the chart. Each line of the chart contains two complementary codons, appearing as pairs consecutively, following their descending thermostability. This is indicated on the left side of the chart as melting enthalpies of respective complementary codon pairs, calculated from original data for ribodinucleotide steps (Xia et al. 1998). Codons for valine and aspartate, according to the chart, are formed by transitions in the middle of the very first codons GGC (Gly) and GCC (Ala). All other codon pairs of the chart are formed by mutations in the third redundant codon positions, and by complementary synthesis of the mutated codon. The last six amino acids of the consensus chronology are in full correspondence with the idea of "codon capture" (Osawa et al. 1992): when all codons are already engaged, the new amino acids can be accommodated only by utilizing (capturing) the codons already assigned earlier.

Size of the Earliest Encoded Peptides

Complementarity principle in the evolution of the triplet code had been introduced first by Eigen and Schuster (1978). It is based on the original observation that alanine and glycine, the highest yield amino acids of the Miller's mix (Miller 1953, 1987), are encoded today by the complementary codons, GGC and GCC, respectively. This complementary pair is also thermodynamically the most stable of all possible complementary codon pairs.

From the complementarity principle it follows (Trifonov et al. 2001; Trifonov 2006) that all amino acids can be divided in two families encoded either by n-Purine-n or n-Pyrimidine-n codons, Gly-family (*G*) and Ala-family (*A*), respectively. Every comple-

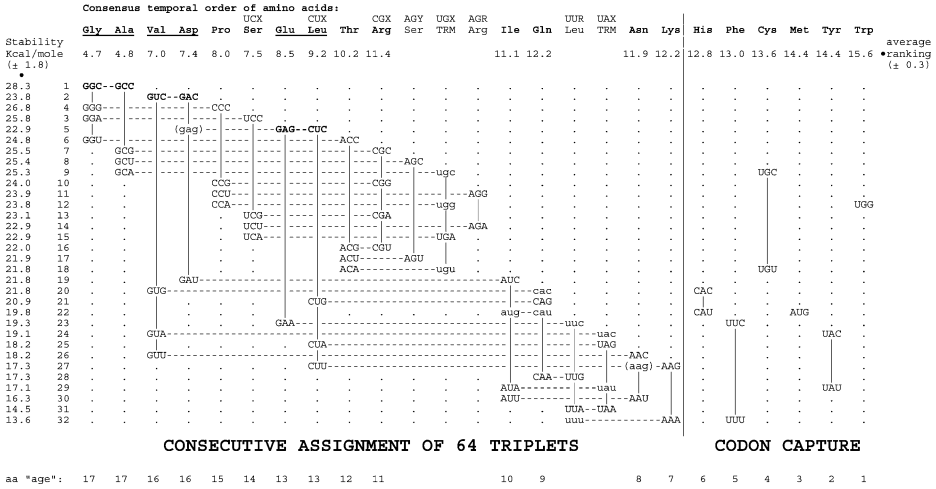

Figure 1 Evolution of the triplet code. The chart is based on the consensus amino acid chronology averaged over 82 different amino acid temporal orders suggested in literature (see the text). For more detailed description of the chart see (Trifonov 2004).

mentary pair of the codons (lines in the Figure 1) corresponds to two different amino acids from the "complementary" families *G* and *A*. In the first step (line) these are Gly(G) and Ala(A), in the second step Asp(D) and Val(V), third – Gly(G) and Pro (P), and so on. The Gly-family: G, D, E, R, S, Q, N, K, H, C, Y and W. The Ala-family: A, V, P, S, L, T, I, F, M. In the course of evolution of the codon table the codons corresponding to Gly family would all appear in the same strand as the original Gly codon GGC, while the Ala family codons – in the complementary strand, that carried the original codon GCC. Such division of the amino acids in two families, presumably, has been strictly maintained during duplex mRNA stage and, perhaps, all the way during evolution of the triplet code and beyond. Indeed, it turned out that amino acid replacements in modern proteins maintain the rule, so that Ala family amino acids are replaced almost exclusively by Ala family residues, and Gly family – by Gly family residues (Trifonov 2006). This confirmation of the fundamental complementarity principle for codons allowed to estimate, from the modern sequences, the size of the most ancient short peptides, at the very onset of the evolutionary history of the codons (Trifonov et al. 2001). For that purpose the sequences were rewritten in the binary *G* and *A* alphabet, and the ancient picture of alternating patches of *A* and *G* residues reconstructed, appearing as well detectable oscillation hidden in the sequences. The very first mini-genes are found to have contained only 6–7 codons, for mini-proteins A_{6-7} and G_{6-7} (Trifonov et al. 2001).

Early Mini-Hairpins of mRNA

These short mini-genes could exist either in a duplex form, two complementary genes together, or as separate single-stranded molecules. In this case they could form complementary hairpin-like structures themselves, which would be more stable and, thus, advantageous. Could these hairpins of 18–21 bases, with 8–9 base pair stems be detected as molecular fossils in modern mRNA sequences? Massive mutational changes accumulated

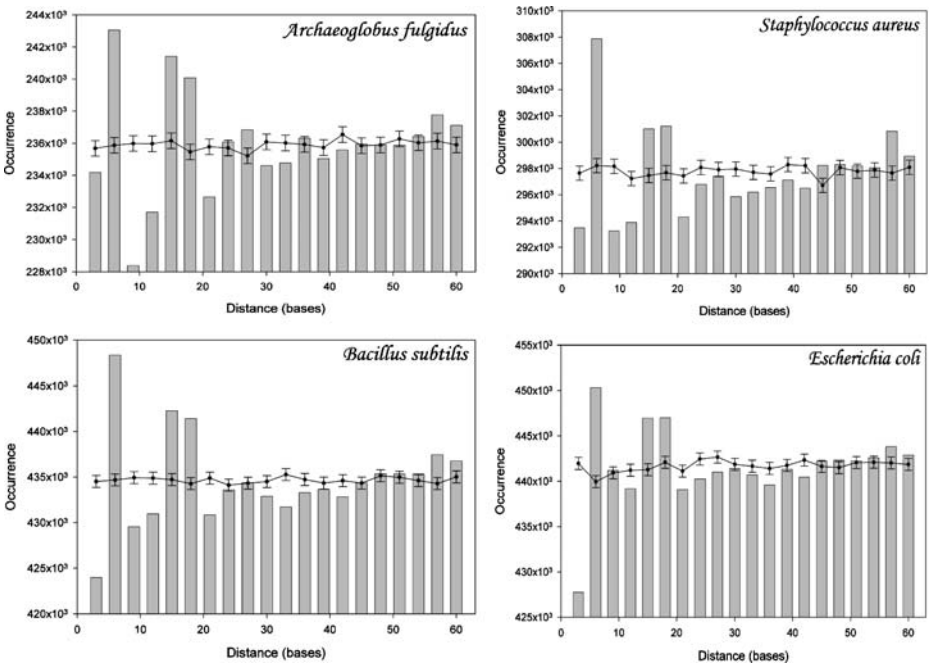

Figure 2 Distance histograms for purine/pyrimidine and pyrimidine/purine contacts in mRNA of four different prokaryotes. The histograms for respective shuffled sequences are shown for comparison, with the error bars, estimated as square roots of the occurrences.

during billions of years leave no hope for survival of the hairpins in their intact original form. Considering, however, relative conservation in amino acid replacements (independent *A* and *G* families) one may expect partial survival of the original complementary middle-codon positions as purines "complementary" to pyrimidines. In this case one would expect that in mRNA sequences the distances between the middle-position purines and pyrimidines in the sequences could still reflect the complementary hairpin structure of the ancient mini-genes. In case of the 9-basepair stem hairpins with 3-base loops the distances 18, 12 and six bases may still appear as more frequent in the histograms of the complementary distances in the modern mRNA sequences. The hairpins of 18 bases (6-basepair stems and 6-base loops) would generate, similarly, the distances 15 and nine bases. That is, the histograms should contain a very characteristic feature – sharp drop in the occurrence of the complementary distances at 18 bases (for the mini-genes encoding 6–7 amino acids). As the Figure 2 demonstrates, indeed, the expected change is observed, exactly at 18 bases. This confirms the size of 6–7 amino acids for the earliest encoded peptides, as estimated above from hidden alternation of patches of *A*- and *G*-families of amino acids in modern proteins.

The Oldest Peptides Extracted from Modern Protein Sequences

One way to detect the peptides is to locate in the protein sequences sections consisting of old amino acids. The age of a given amino acid may be described half-quantitatively by the

number of steps in the consensus amino acid chronology (Trifonov 2004; Sobolevsky and Trifonov 2005), counting from the youngest, last amino acid, tryptophane, of age 1 (see Figure 1). An alternative way to evaluate the age of a given sequence fragment is to measure its degree of conservation. The most conserved sequences are omnipresent, that is found in all known proteomes. The degree of conservation may be measured by the proportion of species harboring the sequence fragment. The older the sequence, the more conserved it is, i.e., the more proteomes contain the fragment. As it has been recently shown (Sobolevsky and Trifonov 2005), the degree of omnipresence strongly correlates with the evolutionary age of the conserved sequences, as calculated from the numbers of steps in the amino acid chronology.

The Functions of the Earliest Proteins

Modern functional involvement of the omnipresent peptides would not necessarily correspond exactly to their original functions. It is natural, however, to expect that in whatever capacity they excelled in the past, that capacity is still their strong characteristic today. If, for example, the motif LSGGQQQR, a typical element of ABC transporters today, plays some particular role in the transporters, its original role has been, likely, the same when the element acted in much simpler molecular environment, perhaps, as part of a very small molecule. The hypothetical small molecule could have been ancestral to the ABC transporters, or one of the modules of which the transporters are built. The modular structure of modern proteins is suggested by sequence conservation plots where the numerous highly conserved short motifs are indicated by sharp peaks at various sequence positions (Aharonovsky and Trifonov 2005).

On the Abiotic Component in the Earliest Life

The Table I lists those modern proteins where the highly conserved elements appear. These octamers are not omnipresent, except for the one on the top. This would mean that the octamers suffered some changes at various stages of evolution. Inspection of the list of respective protein functions reveals that the top conserved octamers are not involved in elementary syntheses. Only in the lines 28 and 31 the first activities of this kind appear (enolase and CTP synthase), followed by other enzymes involved in the synthesis of RNA monomers.

The primacy of RNA in early evolution has long been an accepted view. But where are the enzymes synthesizing amino acids? Inspection of the full list of the octamers (not shown) suggests that the first amino acid enzymes appeared only after several hundred steps (conserved in about 70% genomes). Reductase that reduces ribose to deoxyribose appears at the conservation level <50% of genomes. Thus, enzymatic production of RNA components appears to have been first, amino acid syntheses next, and synthesis of DNA monomers – last. Essentially, it is very close to what one would expect a priori. What is completely counterintuitive is seemingly full life, with replication, transcription and translation, but without any internal material resource at the earliest conserved stages of molecular evolution, as it is witnessed by the oldest omnipresent peptides. The simple suggestion bursts in: all the elementary syntheses in the beginning occurred abiotically, and the monomers had been supplied by environment. Wild as it sounds, the suggestion is not without good independent support. First, this is of course classical synthesis of amino acids

TABLE I Functional involvement of the most conserved octamers (Sobolevsky and Trifonov 2005), present in all (131) or almost all (125 and less) prokaryotic proteomes

Octamer	Present	Prokaryotic proteomes
GHVDHGKT	131	Initiation and elongation factors
SGSGKSTL	125	ABC transporter family proteins
LSGGQQQR	125	ABC cassettes
GPPGTGKT	122	Cell division proteins
KMSKSLGN	121	aa-tRNA synthetases class I
LRPGRFDR	119	Cell division proteins
QRVAIARA	119	ABC cassettes
DEPTSALD	119	ABC cassettes
SIGEPGTQ	117	DNA-directed RNA polymerases
SGGLHGVG	117	Topoisomerases
VEGDSAGG	116	Topoisomerases
GLPNVGKS	116	GTP/ATP binding proteins
DEPSIGLH	115	Exinuclease ABC (UvrA)
DLGGGTFD	115	Chaperones (heat shock) proteins
GPNGAGKS	114	ABC transporters
GIDLGTTN	113	Chaperones
VITVPAYF	113	ATPase of heat shock protein 70
LNRAPTLH	113	RNA polymerase beta' subunit
NADFDGDQ	113	RNA polymerase beta' subunit
NLLGKRVD	113	RNA polymerase beta' subunit
AGDGTTTA	112	Chaperonin GroEL
GPTGVGKT	112	Chaperone ClpB
GIAVGMAT	112	DNA gyrase subunit A
GFDYLRDN	112	Preprotein translocase secA subunit
ERERGITI	111	GTP-binding protein lepA
KPNSALRK	111	30S ribosomal protein S12
NMITGAAQ	111	Elongation factor TU
SHRSGETE	110	**Enolase (phosphopyruvate hydratase)**
MAGRGTDI	110	Preprotein translocase secA subunit
IIFIDEID	110	cell division protein FtsH
GGTVGDIE	110	**CTP synthase**
KFSTYATW	109	RNA polymerase sigma factor rpoD
DEARTPLI	108	Preprotein translocase secA subunit
HHNVGGLP	108	**GMP synthase [glutamine-hydrolyzing]**
GHNLQEHS	107	30S ribosomal protein S12
GGRVKDLP	107	30S ribosomal protein S12
LPDKAIDL	107	Chaperone ClpB
NPRSTVGT	107	Excinuclease ABC subunit A
NEKRMLQE	106	DNA-directed RNA polymerase beta' chain
CPIETPEG	106	DNA-directed RNA polymerase beta chain
NPETVSTD	106	**Carbamoyl-phosphate synthase large chain**
LEYRGYDS	106	**Glucosamine–fructose-6-phosphate aminotransferase**
SRSSALAS	106	**Carbamoyl-phosphate synthase large chain**
HTRWATHG	106	**Glucosamine–fructose-6-phosphate aminotransferase**
DEREQTLN	105	Cell division protein FtsH
DVSGEGVQ	105	ATP-dependent Clp protease ATP-binding subunit clpX
GPSGCGKS	105	Phosphate import ATP-binding protein pstB
KTKPTQHS	105	**CTP synthase**
DHPHGGGE	105	50S ribosomal protein L2
GRFRQNLL	105	DNA-directed RNA polymerase beta' chain

in the imitation experiments of Miller (1953, 1987). Nine of 10 amino acids of the Miller's mix appear at the top of the consensus chronology of amino acids (Figure 1). That is, with these 10 amino acids the Life could, indeed, proceed from its onset to the template syntheses without any domestic supply of the amino acids. As the Table I suggests, the first elementary building elements not readily supplied by the environment, had been nucleobases and ribonucleotides. In its earliest stages Life, apparently, could do well with the small amounts of the RNA monomers generated abiotically. One system of naturally catalyzed synthesis of cytosine and purine in appreciable amounts is TiO_2 catalysis in presence of formamide (Saladino et al. 2004). It is, thus, possible to make at least some parts of RNA abiotically. The sequence analysis data presented in this paper is an encouragement for further studies on the abiotic syntheses, which may well turn to be more generous and diverse than it is currently believed.

References

Aharonovsky E, Trifonov EN (2005) Protein sequence modules. J Biomol Struct Dyn 23:237–242

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

Miller SL (1953) A production of amino acids under possible primitive Earth conditions. Science 117:528–529

Miller SL (1987) Which organic compounds could have occurred on the prebiotic Earth? Cold Spring Harbor Symp Quant Biol 52:17–27

Osawa S, Jukes TS, Watanabe K, Muto A (1992) Recent evidence for evolution of the genetic code. Microbiol Rev 56:229–264

Saladino R, Crestini C, Costanzo G, DiMauro E (2004) Advances in the prebiotic synthesis of nucleic acids bases: implications for the origin of Life. Curr Org Chem 8:1425–1443

Sobolevsky Y, Trifonov EN (2005) Conserved sequences of prokaryotic proteomes and their compositional age. J Mol Evol 61:591–596

Trifonov EN (2000) Consensus temporal order of amino acids and evolution of the triplet code. Gene 261: 139–151

Trifonov EN (2004) The triplet code from first principles. J Biomol Struct Dyn 22:1–11

Trifonov EN (2006) Theory of early molecular evolution: predictions and confirmations. In: Eisenhaber F (ed) Discovering biomolecular mechanisms with computational biology. Landes Bioscience, Georgetown, p 107–116

Trifonov EN, Kirzhner A, Kirzhner VM, Berezovsky IN (2001) Distinct stages of protein evolution as suggested by protein sequence analysis. J Mol Evol 53:394–401

Xia T, SantaLucia J, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH (1998) Thermodynamical parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 37:14719–14735