# EARLY MOLECULAR EVOLUTION

Edward N. Trifonov

Genome diversity Center, Institute of Evolution, University of Haifa, Mount Carmel, 31905 Haifa, Israel

**Israel Journal of Ecology and Evolution, in press**

Running title: Early molecular evolution.

Contact:

Edward N. Trifonov

Genome diversity Center, Institute of Evolution, University of Haifa, Mount Carmel, 31905 Haifa, Israel

phone 04 828 8096, fax 04 824 6554

trifonov@research.haifa.ac.il

## Abstract

Four fundamentally novel, recent developments make a basis for the Theory of Early Molecular Evolution. The theory outlines the molecular events from the onset of the triplet code to the formation of the earliest sequence/structure/function modules of proteins. These developments are: 1) Reconstruction of the evolutionary chart of codons; 2) Discovery of omnipresent protein sequence motifs, apparently conserved since the last common ancestor; 3) Discovery of closed loops - standard structural modules of modern proteins; 4) Construction of protein sequence space of module size fragments, with far reaching evolutionary implications. The theory generates numerous predictions, confirmed by massive nucleotide and protein sequence analyses, such as existence of two distinct classes of amino acids, and their periodical distribution along the sequences . The emerging picture of the earliest molecular evolutionary events is outlined - consecutive engagement of codons, formation of the earliest short peptides, and growth of the polypeptide chains to the size of loop closure, 25-30 residues.

**Keywords:**

Early molecular evolution, evolutionary chart of codons, evolution of code, early proteins, abiotic start, closed loops, omnipresent protein sequence motifs, last common ancestor

## Introduction

Development of living matter from simple to complex forms is in the very basis of evolutionary theory, and numerous studies have been devoted to reconstruction of hypothetical early forms, the last universal common ancestor (LUCA), in particular (e. g., Mirkin et al., 2003). One could also go deeper in the past, beyond the LUCA, by asking questions about even earlier purely molecular events. What were the very first living molecules like? What were the sequences of the earliest proteins and nucleic acids? These are issues of the study of Early Molecular Evolution, a rather modestly populated field of science pioneered by S. Miller (1953, 1987),  M. Eigen, and colleagues (Eigen and Schuster,1978; Eigen and Winkler-Oswatitsch, 1981a,b).

The author of this review is involved in the studies on early molecular life since 1987. Several rather exciting developments, all targeted at the earliest steps of Life, from first codons to the first small proteins and RNAs, are outlined.

## Reconstruction of the origin and evolution of the triplet code

The reconstruction became possible after the temporal order of appearance of amino acids in evolution was  determined as the consensus of more than 80 different expert opinions, physico-chemical, and biological views (Trifonov, 2000; 2004; Trifonov *et al.*, in press). The consensus order (in one-letter amino acid presentation) is: A/G (the earliest), D/V, P, S, E/L, T, R, I, Q, N, K, H, F, C, M,Y, W (the latest). It is rather robust as it does not change with the addition of new criteria. On the basis of this chronology, the succession of steps in the development of the triplet code is reconstructed following the most basic principles, in form of  the evolutionary chart of codons (Figure 1). As described in details in (Trifonov, 2004), the principles on which the chart is based are 1) *Abiotically synthesized amino acids come first*. These are, as the consensus chronology shows, those amino acids that have been demonstrated to form in conditions imitating primordial atmosphere of Earth (Miller, 1953; 1987); 2) *Complementarity*. The 64 codons appeared as complementary pairs, in 32 steps; 3) *Thermostability*. Most stable codon pairs are assigned before less stable ones; 4) *Processivity*. New codons were point mutated versions of  already engaged codons, and their complements; 5) *Codon capture last*. The latest amino acids captured their codons from excessive earlier codon repertoires.

According to the complementarity principle, all codons first appeared as complementary pairs (lines in the chart). This implies that the first protein-coding mRNA molecules existed in the form of duplexes, with both strands coding. The thermostability principle is reflected in the consecutive engagement of less stable complementary codon pairs, starting with a GGC•GCC pair (for glycine and alanine) and ending with AAA•UUU (for lysine and phenylalanine). From the resulting chart, several important predictions follow that are confirmed by computational sequence analyses. One such prediction is that more ancient proteins would be expected to be glycine-rich (see first 6 lines of the evolutionary chart of codons). Composition analysis of shared sequence segments in prokaryotic and eukaryotic protein homologues confirms this prediction. In fact, a "glycine clock" provides a reconstruction of a *rooted* evolutionary tree for major kingdoms that is consistent with current knowledge (Trifonov, 1999).

The very first amino acids of the triplet code history, alanine (A) and glycine (G), were encoded by complementary GCC and GGC codons, as is the case today. In all further steps, except the second one (transition in the middle position of GGC and GCC codons from G and/or C to A and/or U, respectively), new codons appeared via mutation in the degenerate (redundant) third position, and subsequent complementary copying (see Figure 1). In this process, the middle bases did not change. The 20 letters of the protein sequence alphabet can be, thus, divided into two families, with purine-central codons (*Gly* family, *G*): G (GGN), D (GAY), E (GAR), R (CGN and AGR), S (AGY), Q (CAR), N (AAY), K (AAR), H (CAY), C (UGY), Y (UAY) and W (UGG), chronologically, and pyrimidine-central codons (*Ala* family, *A*): A (GCN), V (GUN), P (CCN), S (UCN), L(CUN and UUR), T (ACN), I (AUY and AUA), F (UUY) and M (AUG). During the formation of the triplet code each of the families evolved by substitutions of old codons NRN (NYN) by new codons from within the same family.

Modern protein sequences can be written in binary form (letters *A* and *G*), thus, suggesting how the most ancient ancestor of the given sequence, of original Ala and Gly residues only (first line of the chart), may have looked. The hypothetical first short peptides, $G_n$ (where n refers to sequence length) and $A_n$ could not be longer than 7-8 residues, as both alanine and glycine are non-polar, and the solubility of short $A_n$ and $G_n$ peptides is only due to the charges at the ends of the peptides (Ogata *et al*., 2000). The earliest short "proteins" of *A*-type and *G*-type amino acids should have had the structure $A_n$ and $G_n$ that may still be reflected in modern protein sequences. Proteins of the next stage would be expected to consist of a mosaic of $A_n$ and $G_n$ segments. The *A* family amino acids are largely non-polar, while *G* family amino acids are polar. Since long chains consisting of only polar or only non-polar residues are either structureless or insoluble, respectively, the mosaic would be expected to be rather an alternation of short $A_n$ segments with short $G_n$ segments. Indeed, positional correlation analysis of modern protein sequences revealed the expected alternation, though understandably, well hidden (Trifonov *et al.*, 2001). The length n is found to be 6-7 residues, and presumably equals the length of the earliest proteins. This size of the earliest proteins (peptides) is confirmed also by detected traces of ancient hairpins in mRNA (Gabdank *et al.*, 2006; Trifonov *et al.*, in press).

An important prediction that follows from the reconstructed evolutionary chart is conservation of middle purines and middle pyrimidines of the codons. This results directly from the basic mechanism of formation of the new codons. The conservation of the middle codon positions may have survived during later stages of evolution, after the codon table had been completed. This expectation is very much confirmed, as shown by analysis of amino acid replacements in modern proteins (Trifonov, 2005; Gabdank *et al.*, 2006).

All these multiple lines of evidence follow directly from the evolutionary chart of codons (Trifonov, 2004), which thus serves as an initial basis for reconstruction of the earliest molecular life. Additional details of intriguing early stages of evolution are outlined below, including results obtained by three other independent approaches.

## Closed loops - modules of protein structure

According to recent studies initiated by the seminal paper of Berezovsky et al.(2000), globular proteins are universally built from nearly standard size units - closed loops of 25-30 amino acid residues. The early proteins at some stage, presumably, appeared as individual closed loops, assembled in later proteins as consecutively connected elements (Trifonov et al., 2001; Trifonov and Berezovsky, 2003). When the protein structure is viewed, the closed loops are not immediately recognized as the building units of the protein. Yet, it is the most natural element of every folded polymer chain, proteins included. When freely immersed in solution, the polymer chains adopt various trajectories induced by (random) thermal motion, due to the flexibility of the chain. Occasionally the chain path comes close to itself, making a loop (ring) of the contour length dependent on the flexibility of the chain (Shimada and Yamakawa, 1984). The optimal contour of polypeptide chains with mixed sequences, as in proteins, is 20 to 40 amino acid residues (Berezovsky *et al*., 2000). Natural protein chains should, of course, follow the physics of polymers, and apparently, they do since the most frequent size of the closed loops in crystallized proteins is 25-30 residues (Berezovsky *et al*., 2000). The closed loops make up the proteins like petals form flowers (Berezovsky and Trifonov, 2002). Naturally, the formation of the closed loops should have been an inevitable stage in protein evolution. Their closure would have been facilitated by the interactions between the residues at the ends of the loops. Thus, the respective sequence biases would be expected in the natural sequences that should have appeared in evolution to accommodate this polymer-statistical feature (Trifonov and Berezovsky, 2003). Such bias, indeed, has been discovered. Namely, the hydrophobic residues in prokaryotic proteins have been found to be preferentially arranged in clusters, 25-30 residues from one another along the sequences (Berezovsky *et al*., 2001; Trifonov et al., 2001, Aharonovsky and Trifonov, 2005a). The contacting hydrophobic clusters at the ends of the closed loops form van der Waals "locks", so that the loops can be called  "loop-n-lock" structures (Berezovsky and Trifonov, 2001; Aharonovsky and Trifonov, 2005a).

The search for hypothetical ancestral prototype sequences from which many modern sequences of the closed loops presumably originated, resulted in a small starting set of seven sequences PI to PVII (Berezovsky et al., 2003a, 2003b), also named by Hebrew letters (Berezovsky and Trifonov, 2002). The most prominent ones are prototypes Aleph (PI) and Beth (PII), with their distinct secondary structures and functional involvements (ATP/GTP binding and ATPases, respectively).

The fundamental significance of the discovery of the closed loops/modules is as yet little appreciated by the protein research community. There is no doubt, however, that this transparent and revealing picture of the protein structure will soon take over, and the protein sequence/structure/function studies will follow the lead.

**Protein sequence fossils.**

Vestiges of the Last Universal Common Ancestor (LUCA) can be found in extant proteins in the form of entirely conserved short sequences present in all, or almost all, sequenced prokaryotic proteomes (*omnipresent* motifs) (Sobolevsky and Trifonov, 2005). One could think of the universal presence of these elements as the result of a horizontal transfer at some evolutionary stage. Thus, the stage would have to be very early to ensure their presence in all thoroughly diverse prokaryotic genomes, Eubacterial and Archaeal alike, hence, the obvious connection with LUCA (Last Universal Common Ancestor). The most conserved motifs found have a size of 7-9 residues. They are listed in Table 1 (Y. Sobolevsky, pers. communication).

Two distinct families can be singled out from the list: one with universal GKT/S element represents Walker A motifs (Walker *et al.*, 1982). These belong to the closed loop prototype Aleph (Berezovsky *et al.*, 2003a, Berezovsky and Trifonov, 2002, Sobolevsky and Trifonov, 2006), responsible for ATP/GTP binding. The second family represents elements of the consensus of the prototype Beth (Berezovsky *et al.*, 2003a, Berezovsky and Trifonov, 2002, Sobolevsky and Trifonov, 2006), involved in ATPases.

The search for the occurrence of the omnipresent motifs in crystallized proteins revealed that practically in all cases each motif is found within closed loops of identical structures conserved within the 25-30 residue span, and the structures belong to the same functional type of proteins. In other words, the omnipresent elements represent the signatures of various, most conserved, closed-loop modules (Sobolevsky and Trifonov, 2006).

Less conserved modules can be detected as well by use of so-called conservation plots (Aharonovsky and Trifonov, 2005b). For every, say, 7-letter long segment of the protein sequence of interest, the number of occurrences of this "word" in all prokaryotic proteomes can be counted. The maxima in the resulting "skyline" plot indicate locations of the signatures of various modules along the protein sequence. Such analysis reveals that the sequence modules follow one another at the distance 25 - 30 residues, in agreement with the view that globular proteins consist of modules (closed loops) of this size. The conservation plots indicate sequence positions of the closed loops in the proteins, including those for which their respective structures are not yet established. That is, they give initial information on the arrangement of the closed loops along the molecule which is further folded in the final structure.

## Evolutionary walks in protein sequence space

In view of the modular structure of proteins, commonly performed whole-length comparisons of the protein sequences do not make sense, except in cases when the proteins are built from the same modules, and thus, are similar sequence-wise all throughout. Accordingly, the evolution of proteins is rather the evolution of their modules and combinations thereof. In other words, the relatedness of the proteins is reduced to the relatedness of the shared modules. While closely related proteins (of the same modular structure and composition) can be still compared by traditional techniques, the evolution of the modules should be traced in the space of short module-sized (or shorter) sequences derived by fragmentation of all available proteins.

Such sequence space has been recently generated from the 20 residue long fragments of all proteins of complete proteomes of 112 prokaryotes (Frenkel and Trifonov, 2006). Connecting the points (fragments) of the sequence space that are sequence-wise similar beyond a given threshold, one generates the sequence "walks" and whole networks of thus related fragments. An example of such network (in this case - fragments from ABC transporter proteins) is shown in the Figure 2. A rather unexpected result (Frenkel and Trifonov, 2006; in press) is that if one walks from one sequence fragment to a closely related one, and farther on, comparing only immediate neighbors along the "walk", the fragments at the ends of the walks almost completely loose any similarity. Yet, typically all the fragments of the walk belong to the same protein type and function. At the same time, transitions are frequently observed such that two adjacent (very similar) fragments in the walk belong to rather different protein types.

The walks calculated for any given threshold of similarity form networks each of the many thousands of the fragments. Some of the networks are linked to one another by one or several bonds (closely related pairs of sequence fragments), indicating that there is likely an evolutionary connection between the networks.

Our attempt to construct "evolutionary trees" in such a protein-sequence space yielded exciting results. To begin with, at the level 10-15% of sequence similarity (simple match), *all* sequence fragments, about 90,000,000 total, combine in one thick trunk of the tree. This, however, is trivial since with such a low threshold even random sequence space would make one big network. At a level about 65% of sequence identity, two distinct bundles of branches are formed. They are found to represent those two large families mentioned before - close relatives of prototype Aleph, and close relatives of prototype Beth (data not shown).

### The earliest protein functions

Presence of a given sequence motif in all or almost all proteomes is an indication that the motif is evolutionarily old. It is unlikely that such omnipresence is the result of horizontal transfer, unless it happened at some early stage, with a very small number of species existing at that time. In other words, the omnipresent motifs are likely to represent the Last Universal Common Ancestor (LUCA). Some of the most ancient motifs could well have disappeared, perhaps, together with their functions and left no traces in modern sequences. Thus, one cannot be sure that reconstruction does not miss something of crucial value for the reassembled picture of the past. Still, even a partial, incomplete answer to the question of what were the most ancient functions of the proteins is of burning interest. Such a partial answer is given by the list of functions associated with the conserved octa-peptides, from the earliest omnipresent and very frequently encountered to the more recent, less frequently encountered sequences. Such a list of over 2000 octamers is calculated in (Sobolevsky and Trifonov, 2005) on the basis of 131 prokaryotic genomes/proteomes, and the respective earliest functions are outlined (Trifonov *et al.*, in press). Table 2, adapted from this work, is shown below. It lists the most ancient protein functions and respective octamers associated with the proteins. Noteworthy is that the search for omnipresent motifs in larger number of proteomes should result in shorter motifs. Table 1 lists the omnipresent motifs of various lengths (6 to 9 residues) calculated from a selection of only 15 phylogenetically diverse proteomes. The omnipresent nonamers of the list, obviously, do not appear in some of the 131 proteomes contributing to Table 2 though their various single-letter mutations are found in every proteome.

What can one say *a priori* about the most ancient functions? Since biological macromolecules had to be synthesized from corresponding monomer units, amino acids, nucleotides and other elementary building blocks, one would expect that the proteins (enzymes) involved in their biosyntheses would be amongst the oldest. Table 2 demonstrates that as far as amino acids are concerned, this is not the case. Apparently, there was no early need for the synthesis of amino acids. Indeed, the earliest amino acids were most probably synthesized abiotically, as originally suggested by the experiments of Miller (1987). The "chronological" list of protein functions in Table 2, where the oldest (harbored by larger number of genomes) octamers are on the top, also suggests that perhaps all other basic elementary building units had been around as well (generous abiotic start). The primary substance in need should have been ATP, a universal energy source for almost all cellular syntheses. Consistent with this is the observation that the oldest, most conserved sequences at the top of Table 2 are associated with ATPases (Table 2, squares) and ATP/GTP binding activities (bullets), and most of the sequences with ATPase and ATP/GTP binding activities are highly conserved. Though initial small amounts of ATP could be also provided abiotically, the emerging living system would, certainly, need a more substantial supply of ATP, i. e., the synthesis of ATP by ATP synthases. Perhaps, the ATPases could have served in this capacity, due to their reverse activity. The first enzyme involved in elementary syntheses other than ATP synthesis appears in Table 2 only at step 28 (enolase) and yet still takes part in the synthesis of ATP. Other elementary syntheses, further on, are indicated in bold. The first elementary synthesis had to be introduced, of course, to supplement the limited supply of some simple molecules from the environment. Activities involved in the syntheses of the amino acids, deoxyribose and thymine appear well below the short list in Table 2. This confirms the commonly accepted idea that DNA appeared only after prior development of the protein/RNA world. Activities related to DNA appear in the list as early as in step 9 (Table 2). This should not be taken, however, as evidence in favor of the early presence of DNA. Respective functions could well be originally geared to RNA.

At the very top of the list of omnipresent motifs (Table 1) one finds the longest sequences, nonamers HVDHGKTTL (family Aleph) and LSGGQQQRV and QRVAIARAL (both family Beth). The last two can be combined in a longer sequence LSGGQQQRVAIARAL, while the first one can be extended to GHVDHGKTTLL, on the basis of a consensus sequence around the omnipresent nonamer, calculated from all the sequences where it is found. These segments (LSGGQQQRVAIARAL and

GHVDHGKTTLL) are the likeliest representatives of the original ancient prototypes Beth and Aleph, respectively. In their earliest form, in binary presentation, they would appear as *AAGG**GGGGAAAAGAA*** and ***GGAGGGGAAAA*.** The RNA sequences encoding the oligopeptides in bold would be complementary to each other, via conserved purines and pyrimidines in the middle of respective triplets (see the section on the chart of codons). Indeed, GGC for *Gly* is complementary to GCC for *Ala*, and ***A's*** are "complementary" to ***G's*** (one of the undecamers has to be read in the opposite direction, to reflect different polarities of the strands in the RNA duplex). This striking complementary pair in bold is a single identified case so far, and their complementarity may be coincidental. It is tempting, however, to speculate that the very first fundamental pedigree genes responsible for ATP-based energy consumption originated, indeed, from different strands of the same RNA duplex. After all, these are the most ancient genes, by their omnipresence and their weight in the evolutionary tree of the sequence space, and last, but not least - due to undeniable primacy of energy in Life.

No speculations were needed to bridge various stages of the reconstruction described above; none of the stages invoked any complications. To begin with, it appears that the codon assignments, once established, did not change anymore. One should have enough courage to make an *a priori* statement like this. Yet, it naturally follows from the reconstruction. The consecutive build up of the evolutionary stages of the triplet code did not require assuming that some other amino acids may have been utilized in the earliest billennia of Life. This does not exclude, of course, some unusual assignments and non-canonical amino acids at some stage. The main stream evolution, apparently, outweighed these possible side trips. Interestingly, recently discovered non-canonical amino acids selenocysteine (Chambers et al., 1986) and pyrrolysine (Srinivasan et al., 2002) are both encoded by triplets captured from termination codons (UGA and UAG, respectively). This behavior is typical for all late amino acids of the codon capture stage. The earliest peptides turned out to be the size suggested by the solubility limit for $A_n$ and $G_n$ (6-7 amino acid residues). Ancient small mRNA hairpins turned out to correspond to the same size (18-21 bases encoding 6-7 amino acids). For the growing polypeptide chains, the first convolution was also the simplest possible - formation of the closed loops of standard size, as dictated by polymer statistics. Finally, inspection of the functions of the proteins harboring omnipresent (most ancient) sequence motifs points to exactly those functions that common sense would suggest. That is, the top of the list does not contain enzymes responsible for synthesis of the earliest amino acids (those that are found in the imitation experiments of Miller). Apparently, there was no need for synthesis of something that was available anyway. The list starts with template functions, cell division, transporters, etc., but not with elementary syntheses. It appears that all other elementary building units had been available as well, having been synthesized abiotically. This almost vulgar opportunism itself is the simplest rule one could think of for the emergence of Life, and as a matter of fact, for all subsequent diversification of Life, including the opportunism of human behavior.

The simplicity of the reconstruction is, undoubtedly, its merit. It confers credibility to the picture. Fortified by the confirmations of several important predictions, suggested by the picture, the reconstruction of the earliest molecular stages of Life appears to make a basis for the vibrant, quickly developing Theory of Early Molecular Evolution.

## Suggestions to experimentalists.

The above theory is internally consistent to such a degree, that fast development of the theory by "dry" means is more rewarding than potentially slow experimental progress. In the long run, however, there is nothing more crucial for any theory than solid experimental confirmation. Hence, I would strongly encourage ambitious experimental researchers to participate in the excitement of the development.

An obvious endeavor is to try to imitate the *first* steps of protein-coding Life - to manipulate the first $(GGC)_n$ and $(GCC)_n$ genes and $Ala_n$ and $Gly_n$ peptides, perhaps with the addition of some natural catalyzers (clays and minerals). Another interesting route is to manipulate the prototype and other closed loop modules, trying to design enzymes for biotechnological needs. The establishment of the modular structure of proteins opens new perspectives for functional characterization of proteins and the decomposition of commonly known protein functions into the elementary sub-functions of the modules. This would allow the design of proteins by exploring only functionally crucial modules.

With the understanding of the modular structure of proteins, and with available maps of protein sequence/structure modules (Aharonovsky and Trifonov, 2005b), the problem of protein folding loses much of its complexity. This application beckons the participation of interested parties.

**Acknowledgements.**

Valuable comments of Dr. Z. M. Frenkel are highly appreciated. The work was supported in part by EU grant QLG2-CT-2002-01298.

**References:**

Aharonovsky, E., Trifonov, E. N. 2005a. Sequence structure of van der Waals locks in proteins. J. Biomol. Str. Dyn **22**: 545-554.

Aharonovsky, E., Trifonov, E. N. 2005b. Protein sequence modules. J. Biomol. Str. Dyn. **23**: 237-242.

Berezovsky, I. N., Trifonov, E. N. 2001. Van der Waals locks: loop-n-lock structure of globular proteins. J. Molec. Biol. **307**: 1419-1426.

Berezovsky, I. N., Trifonov, E. N. 2002. Flowering buds of globular proteins: transpiring simplicity of protein organization, Comp. Funct. Genom. **3**: 525-534.

Berezovsky, I. N., Grosberg, A. Y., Trifonov, E. N. 2000. Closed loops of nearly standard size: common basic element of protein structure. FEBS Letters **466**: 283-286.

Berezovsky, I. N., Kirzhner, V. M., Kirzhner, A, Trifonov, E. N. 2001. Protein folding: looping from hydrophobic nuclei. Proteins: Structure, Function and Genetics **45**: 346-350.

Berezovsky, I. N., Kirzhner, V. M., Kirzhner, A., Rosenfeld, V. R., Trifonov, E. N. 2003a. Protein sequences yield a proteomic code, J. Biomol. Struct. Dyn. **21**: 317-325.

Berezovsky, I. N., Kirzhner, A., Kirzhner, V. M., Trifonov, E. N. 2003b. Spelling protein structure, J. Biomol. Struct. Dyn. **21**: 327-339.

Chambers, I. J., Frampton, J., Goldfarb, P., Affara, N., McBain, W., Harrison, P. R. 1986. The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active site is encoded by the “termination” codon, TGA. EMBO J. **5**: 1221-1227.

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

Eigen, M., Winkler-Oswatitsch, R. 1981a. Transfer-RNA: The early adaptor. Naturwissenschaften **68**:217-228.

Eigen, M., Winkler-Oswatitsch, R. 1981b. Transfer-RNA, an early gene? Naturwissenschaften **68**:282-292.

Frenkel, Z. M., Trifonov, E. N. 2006. Walking through protein sequence space. J. Theor. Biol. **244**: 77-80.

Frenkel, Z. M., Trifonov, E. N. Walking through the protein sequence space: Towards new generation of the homology modeling. Proteins Str. Function Bioinf., in press.

Gabdank, I., Barash, D., Trifonov, E. N. 2006. Tracing ancient mRNA hairpins. J. Biomol. Str. Dyn. **24**: 163-170.

Miller, S. L. 1953. A production of amino acids under possible primitive earth conditions. Science **117**:528-529.

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

Mirkin, B. G., Fenner, T. I., Galperin, M. Y., Koonin, E. V. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol. Biol. **3**: 2.

Ogata, Y., Imai, E-I., Honda, H., Hatori, K., Matsuno, K. 2000. Hydrothermal circulation of seawater through hot vents and contribution of interface chemistry to prebiotic synthesis. Origin Life Evol. Biosph. **30**:527-537.

Shimada, J., Yamakawa, H. 1984. Ring-closure probabilities for twisted worm-like chains. Application to DNA. Macromolecules **17**: 689-698.

Sobolevsky, Y., Trifonov, E. N. 2005. Conserved sequences of prokaryotic proteomes and their compositional age. J. Mol. Evol. **61**: 591-596.

Sobolevsky, Y., Trifonov, E. N. 2006. Protein modules conserved since LUCA. J. Mol. Evol. **63**: 622-634.

Srinivasan, G., James, C. M., Krzycki, J. A. 2002. Pyrrolysine encoded by UAG in Archaea: Charging of a UAG-decoding specialized tRNA. Science **296:** 1459-1462.

Trifonov, E. N. 1999. Glycine clock: Eubacteria first, Archaea next, Protoctista, Fungi, Planta and Animalia at last. Gene Therapy Molec. Biol. **4**: 313-322. www.gtmb.org

Trifonov, E. N. 2000. Consensus temporal order of amino acids and evolution of the triplet code. Gene **261**: 139-151.

Trifonov, E. N. 2004. The triplet code from first principles. J. Biomolec. Str. Dyn. **22**: 1-11.

Trifonov, E. N. 2006. Theory of early molecular evolution: Predictions and confirmations. In: Discovering Biomolecular Mechanisms with Computational Biology (Ed. F. Eisenhaber), Landes Bioscience, Georgetown, pp 107-116.

Trifonov, E. N., Berezovsky, I. N. 2003. Evolutionary aspects of protein structure and folding, Curr. Opinion Struct. Biol. **13**: 110-114.

Trifonov, E. N., Kirzhner, A., Kirzhner, V. M., Berezovsky, I. N. 2001. Distinct stages of protein evolution as suggested by protein sequence analysis. J. Mol. Evol. **53**: 394-401.

Trifonov, E. N., Gabdank, I., Barash, D, Sobolevsky, Y. *Primordia vita*. Deconvolution from modern sequences. Origin Life Evol. Biosph., in press.

Walker, J. E., Saraste, M., Runswick, M.J., Gay, N.J. 1982. Distantly related sequences in the alpha-subunits and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J. **1**: 945-951

.

Figure legends:

Figure 1. Evolutionary chart of codons, the latest update (Trifonov et al., in press). The chart is based on the consensus amino acid chronology (an upper line) averaged over 84 different amino acid temporal orders suggested in literature. Each of the suggested orders ranks the amino acids chronologically, and the averaged ranks are shown in the chart. The three-letter abbreviations correspond (in the order indicated) to amino acids alanine (Ala, A), glycine (Gly, G), valine (Val, V), aspartate (Asp, D), proline (Pro, P), serine (Ser, S), glutamate (Glu, E), leucine (Leu, L), threonine (Thr, T), arginine (Arg, R), isoleucine (Ile, I), glutamine (Gln, Q), asparagine (Asn, N), lysine (Lys, K), histidine (His, H), phenylalanine (Phe, F), cysteine (Cys, C), methionine (Met, M), tyrosine (Tyr, Y), and tryptophan (Trp, W). The lines of the chart correspond to consecutive steps of engagement of the codons. The codons of the last columns of the chart (Codon capture) are those, that have been already engaged. At this last stage they are reassigned for the latest amino acids. For more detailed description of the chart see (Trifonov, 2004).

Figure 2. Network of connections between pair-wise closely similar sequence fragments (points). The network combines all related fragments of  ABC transporters  of 112 proteomes. The connections correspond to 70% sequence identity between the neighboring fragments (courtesy of Z. M. Frenkel).

Table 1. The most ancient (omnipresent) protein sequence motifs.

| Family Aleph | Family Beth | Other unique motifs |
|---|---|---|
| **HVDHGKTTL** | **QRVAIARAL...LADEPT** | **FIDEID** |
| **GPPGTGKT** | **LSGGQQQRV** | |
| **GHVDHGKT** | **TLSGGE** | **IDTPGHV** |
| **GSGKTTLL** | | |
| **GPSGSGK** | | **KMSKSL** |
| **PTGSGKT** | | |
| **NGSGKTT** | | **NADFDGD** |
| **GKSTLLN** | | |
| **SGSGKT** | | **WTTTPWT** |
| **TGSGKS** | | |
| **PGVGKT** | | |
| **PNVGKS** | | |
| **GVGKTT** | | |
| **GTGKTT** | | |
| **DHGKST** | | |
| **GKTTLA** | | |
| **GKTTLV** | | |
| **KSTLLK** | | |

Table 2. Functional involvement of the most conserved octamers present in all (131) or almost all (125 and less) prokaryotic proteomes.

| | | number of genomes | protein function |
|---|---|---|---|
| 1. | GHVDHGKT | 131 | ● ■initiation and elongation factors |
| 2. | SGSGKSTL | 125 | ● ■ABC transporter family proteins |
| 3. | LSGGQQQR | 125 | ● ■ABC cassettes, transporters |
| 4. | GPPGTGKT | 122 | ●cell division proteins |
| 5. | KMSKSLGN | 121 | aa-tRNA synthetases class I |
| 6. | QRVAIARA | 119 | ● ■ABC cassettes, transporters |
| 7. | DEPTSALD | 119 | ● ■ABC cassettes, transporters |
| 8. | LRPGRFDR | 119 | cell division proteins |
| 9. | SIGEPGTQ | 117 | DNA-directed RNA polymerases |
| 10. | SGGLHGVG | 117 | topoisomerases |
| 11. | VEGDSAGG | 116 | topoisomerases |
| 12. | GLPNVGKS | 116 | ●GTP/ATP binding proteins |
| 13. | DEPSIGLH | 115 | ■exinuclease ABC (UvrA) |
| 14. | DLGGGTFD | 115 | chaperones (heat shock) proteins |
| 15. | GPNGAGKS | 114 | ● ■ABC transporters |
| 16. | GIDLGTTN | 113 | chaperones |
| 17. | VITVPAYF | 113 | ■ATPase of heat shock protein 70 |
| 18. | LNRAPTLH | 113 | RNA polymerase beta' subunit |
| 19. | NADFDGDQ | 113 | RNA polymerase beta' subunit |
| 20. | NLLGKRVD | 113 | RNA polymerase beta' subunit |
| 21. | AGDGTTTA | 112 | chaperonin GroEL |
| 22. | GPTGVGKT | 112 | ●chaperone ClpB |
| 23. | GIAVGMAT | 112 | DNA gyrase subunit A |
| 24. | GFDYLRDN | 112 | preprotein translocase secA subunit |
| 25. | ERERGITI | 111 | ●GTP-binding protein lepA |
| 26. | KPNSALRK | 111 | 30S ribosomal protein S12 |
| 27. | NMITGAAQ | 111 | elongation factor TU |
| 28. | SHRSGETE | 110 | **enolase (phosphopyruvate hydratase)** |
| 29. | MAGRGTDI | 110 | preprotein translocase secA subunit |
| 30. | IIFIDEID | 110 | cell division protein FtsH |
| 31. | GGTVGDIE | 110 | **CTP synthase** |
| 32. | KFSTYATW | 109 | RNA polymerase sigma factor rpoD |
| 33. | DEARTPLI | 108 | preprotein translocase secA subunit |
| 34. | HHNVGGLP | 108 | **GMP synthase** |
| 35. | GHNLQEHS | 107 | 30S ribosomal protein S12 |
| 36. | GGRVKDLP | 107 | 30S ribosomal protein S12 |
| 37. | LPDKAIDL | 107 | chaperone ClpB |
| 38. | NPRSTVGT | 107 | ■exinuclease ABC subunit A |
| 39. | NEKRMLQE | 106 | DNA-directed RNA polymerase beta' chain |
| 40. | CPIETPEG | 106 | DNA-directed RNA polymerase beta chain |
| 41. | NPETVSTD | 106 | **carbamoyl-phosphate synthase large chain** |
| 42. | LEYRGYDS | 106 | **glucosamine-fructose-6-phosphate aminotransferase** |
| 43. | SRSSALAS | 106 | **carbamoyl-phosphate synthase large chain** |
| 44. | HTRWATHG | 106 | **glucosamine-fructose-6-phosphate aminotransferase** |
| 45. | DEREQTLN | 105 | cell division protein FtsH |
| 46. | DVSGEGVQ | 105 | ●Clp protease ATP-binding subunit clpX |
| 47. | GPSGCGKS | 105 | ●phosphate import ATP-binding protein pstB |
| 48. | KTKPTQHS | 105 | **CTP synthase** |
| 49. | DHPHGGGE | 105 | 50S ribosomal protein L2 |
| 50. | GRFRQNLL | 105 | DNA-directed RNA polymerase beta' chain |

```
            Consensus temporal order of amino acids:
                                     UCX       CUX       CGX  AGY  UGX  AGR            UUR  UAX          |
            Ala  Gly  Val  Asp  Pro  Ser  Glu  Leu  Thr  Arg  Ser  TRM  Arg  Ile  Gln  Leu  TRM  Asn  Lys| His  Phe  Cys  Met  Tyr  Trp
Stability                                                                                                |                               average
Kcal/mole   4.8  4.9  7.0  7.3  8.1  7.5  8.4  9.2 10.1 11.4                11.1 12.1           11.9 12.3|12.7 13.1 13.4 14.4 14.4 15.6 ←ranking
(± 1.8)                                                                                                  |                               (± 0.3)
  ↓                                                                                                      |
28.3     1  GCC--GGC   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    . |  .    .    .    .    .    .
23.8     2   |    |   GUC--GAC   .    .    .    .    .    .    .    .    .    .    .    .    .    .    . |  .    .    .    .    .    .
26.8     4   |   GGG---|----|---CCC   .    .    .    .    .    .    .    .    .    .    .    .    .    . |  .    .    .    .    .    .
25.8     3   |   GGA---|----|----|---UCC   .    .    .    .    .    .    .    .    .    .    .    .    . |  .    .    .    .    .    .
22.9     5   |    |    |  (gag)--|----|---GAG--CUC   .    .    .    .    .    .    .    .    .    .    . |  .    .    .    .    .    .
24.8     6   |   GGU---|----|----|----|----|----|---ACC   .    .    .    .    .    .    .    .    .    . |  .    .    .    .    .    .
25.5     7  GCG--------|----|----|----|----|----|----|---CGC   .    .    .    .    .    .    .    .    . |  .    .    .    .    .    .
25.4     8  GCU--------|----|----|----|----|----|----|----|---AGC   .    .    .    .    .    .    .    . |  .    .    .    .    .    .
25.3     9  GCA--------|----|----|----|----|----|----|----|----|---ugc   .    .    .    .    .    .    . |  .    .   UGC   .    .    .
24.0    10   .    .    |    |   CCG---|----|----|----|---CGG   |    |    .    .    .    .    .    .    . |  .    .    |    .    .    .
23.9    11   .    .    |    |   CCU---|----|----|----|----|----|----|---AGG   .    .    .    .    .    . |  .    .    |    .    .    .
23.8    12   .    .    |    |   CCA---|----|----|----|----|----|---ugg   |    .    .    .    .    .    . |  .    .    |    .    .   UGG
23.1    13   .    .    |    |    .   UCG---|----|----|---CGA   |    |    |    .    .    .    .    .    . |  .    .    |    .    .    .
22.9    14   .    .    |    |    .   UCU---|----|----|----|----|----|---AGA   .    .    .    .    .    . |  .    .    |    .    .    .
22.9    15   .    .    |    |    .   UCA---|----|----|----|----|---UGA   .    .    .    .    .    .    . |  .    .    |    .    .    .
22.0    16   .    .    |    |    .    .    |    |   ACG--CGU   |    |    .    .    .    .    .    .    . |  .    .    |    .    .    .
21.9    17   .    .    |    |    .    .    |    |   ACU-------AGU   |    .    .    .    .    .    .    . |  .    .    |    .    .    .
21.8    18   .    .    |    |    .    .    |    |   ACA------------ugu   .    .    .    .    .    .    . |  .    .   UGU   .    .    .
21.8    19   .    .    |   GAU-------------|----|----------------------------AUC   .    .    .    .    . |  .    .    .    .    .    .
21.8    20   .    .   GUG------------------|----|-----------------------------|---cac   .    .    .    . | CAC    .    .    .    .    .
20.9    21   .    .    |    .    .    .    |   CUG----------------------------|---CAG   .    .    .    . |  |    .    .    .    .    .
19.8    22   .    .    |    .    .    .    |    |    .    .    .    .    .   aug--cau   .    .    .    . | CAU    .    .   AUG   .    .
19.3    23   .    .    |    .    .    .   GAA---|-----------------------------|----|---uuc   .    .    . |  .   UUC   .    .    .    .
19.1    24   .    .   GUA-----------------------|-----------------------------|----|----|---uac   .    . |  .    |    .    .   UAC   .
18.2    25   .    .    |    .    .    .    .   CUA----------------------------|----|----|---UAG   .    . |  .    |    .    .    |    .
18.2    26   .    .   GUU-----------------------|-----------------------------|----|----|----|---AAC   . |  .    |    .    .    |    .
17.3    27   .    .    .    .    .    .    .   CUU----------------------------|----|----|----|--(aag)-AAG|  .    |    .    .    |    .
17.3    28   .    .    .    .    .    .    .    .    .    .    .    .    .    |   CAA--UUG   |    |    | |  .    |    .    .    |    .
17.1    29   .    .    .    .    .    .    .    .    .    .    .    .    .   AUA--------|---uau   |    | |  .    |    .    .   UAU   .
16.3    30   .    .    .    .    .    .    .    .    .    .    .    .    .   AUU--------|----|---AAU   | |  .    |    .    .    .    .
14.5    31   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .   UUA--UAA        | |  .    |    .    .    .    .
13.6    32   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .   uuu------------AAA|  .   UUU   .    .    .    .
                                                                                                         |
```

**CONSECUTIVE ASSIGNMENT OF 64 TRIPLETS**                    **CODON CAPTURE**

fig. 1

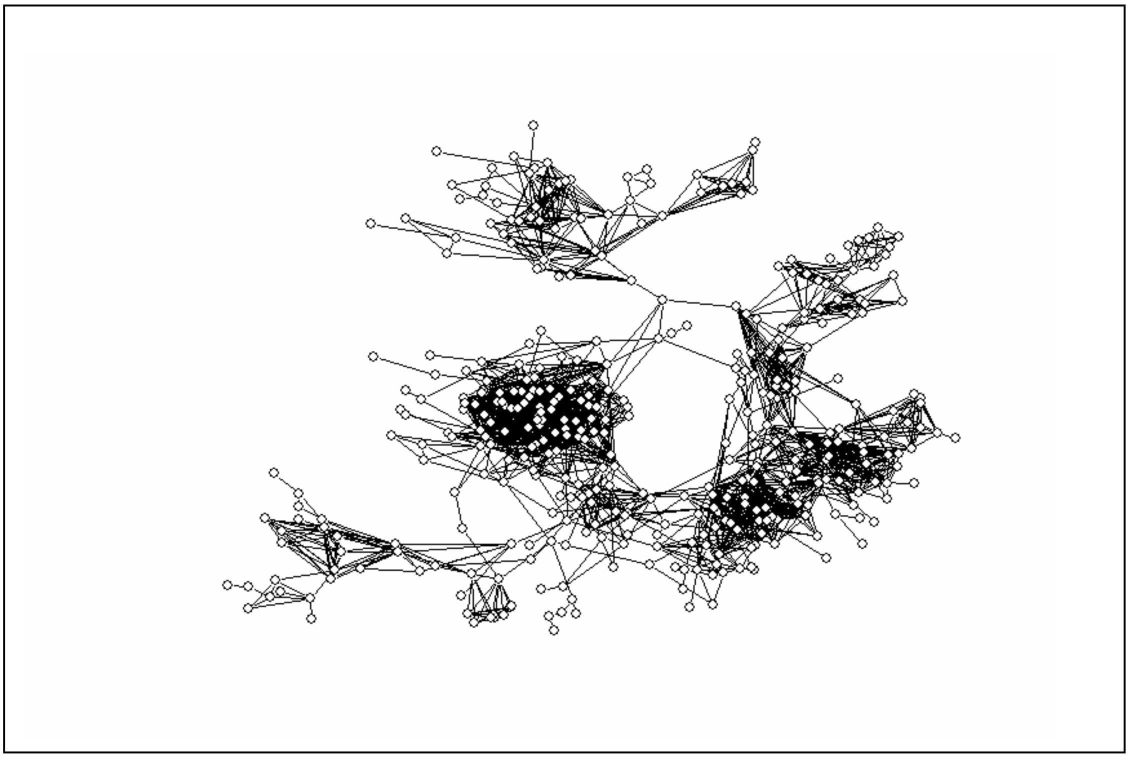

fig. 2