Minireview

Molecular evolution from abiotic scratch

Edward N. Trifonov[a,*], Igor N. Berezovsky[b]

[a] *Genome Diversity Center, Institute of Evolution, University of Haifa, Haifa 31905, Israel*
[b] *Department of Structural biology, The Weizmann Institute of Science, Rehovot 76100, Israel*

Received 11 June 2002; revised 21 July 2002; accepted 23 July 2002

First published online 9 August 2002

Edited by Gianni Cesareni

Abstract Recent papers on the emerging new theory of protein evolution are reviewed. Reconstruction of codon chronology, analysis of loop fold structure of proteins, and quantitative correspondence between optimal DNA ring closure size and protein domain size allow to outline specific stages in early protein evolution, each with its own size range. © 2002 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.

Key words: Protein structure; Origin of the genetic code; Protein evolution; Closed loops; DNA rings

1. Introduction

In the very first eons of life's struggle on the Earth, it was presumably crucial for nucleic acid and protein sequences to develop certain patterns most suitable for their initially simple but vital functions. The patterns adapted during the early stages of molecular evolution may still reside in modern sequences, hopefully in detectable form.

2. Evolution of the genetic code. Reconstruction

The first steps in a search for such ancient patterns were taken by Eigen and Winkler-Oswatitsch [1,2] who were able to reconstruct the earliest hypothetical mRNA/tRNA sequences and to reveal their predicted repeating $(RNY)_n$ pattern. Its presence in modern mRNA sequences was immediately confirmed [3] but it took almost two decades before this line of thought made a fruitful return.

The return was triggered by an observation [4] that so-called triplet expansion diseases – dramatic change in copy numbers of certain tandemly repeating triplets – are expansions of almost exclusively GCT and GCC triplets. It was then already known that the 'consensus' pattern of mRNA is $(G\text{-nonG-}N)_n$ [5] or, in further refined form $(GCU)_n$ [6] rather than $(RNY)_n$. The property to expand during complementary replication should have been a great advantage for the earliest coding sequences. Thus, it was natural to suggest that the GCU (GCC) coding triplets were the earliest [4]. The next codons to come were likely to be single point mutation derivatives of the GCU (GCC) triplets, to code for the earliest amino acids. According to the codon table these must be then Ala, Asp, Gly, Ser, Pro, Val and Thr. Rather straightforward criteria of the amino acid chronology – chemical simplicity, synthesis in the classical imitation experiments by Miller [7], and association with more ancient class II amino-acyl-tRNA synthetases – suggest six of the above seven amino acids as the first to appear on the evolutionary scene [4]. This spectacular result is confirmed by a more extensive analysis of possible amino acid chronology, i.e. by using 44 different criteria of the relative amino acid ages [8,9]. According to the consensus amino acid chronology, the first amino acids of the ancient proteins should have been Gly, Ala, Val, Asp, Pro, Ser, Thr, Glu, and Leu – all from Miller's mixture, which is an important result per se. One may also try to reconstruct the chronology of codons. In Fig. 1 only those 20 codons for 20 amino acids are represented which make the most stable complementary contacts with their counterparts (codon pairs connected by horizontal dotted lines). This partial reconstruction of the codon chronology is largely based on the hypothesis of Eigen and Schuster: that the earliest mRNA existed as duplexes coding in both strands, with alanine and glycine as the very first amino acids encoded, respectively, by the most stable complementary GCC and GGC triplets [10]. Note, that the complementary codons are all on the left side of the diagonal. That is, new codons are derived from the point-mutated versions of the earlier ones as their complementary copies (processivity). Such striking triangularity can only be achieved if the chronological order of the amino acids is, indeed, very close to the derived consensus (upper line of Fig. 1), and if the complementarity and thermostability rules of Eigen and Schuster are respected. The combination of these most natural rules together with the new rule of processivity make the triangular reconstruction in Fig. 1 an attractive basis for further predictions.

The reconstructed codon chronology suggests strong early domination of Gly residues ([8,9], see also Fig. 1). Indeed, four of the first 10 codons (see five first lines in Fig. 1) are the codons for glycine. Therefore the relatively high content of glycine would be predicted in the ancient proteins. Comparison of homologous protein sequences of prokaryotes and eukaryotes allows an estimate of the amino acid composition of proteins at the moment of separation of these large kingdoms, some 3.5 billion years ago, by taking only matching, conserved sections of the homologous sequences [11]. As predicted, the glycine content of these sections is found to be substantially higher (14%) than in modern proteins (7–8%).

[*] Corresponding author. Fax: (972)-4-8246554.
E-mail address: edward.trifonov@weizmann.ac.il (E.N. Trifonov).

PII: S0014-5793(02)03165-4

```
     1    2    3    4    5    6    7    8    9    10   11   12   13   14   15   16   17   18   19   20

    Gly  Ala  Val  Asp  Pro  Ser  Thr  Glu  Leu  Arg  Ile  Gln  Asn  Lys  Cys  Phe  His  Tyr  Met  Trp


 1  GGC..GCC   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
 2   :    :   GUC..GAC   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
 3  GGG...:....:....:...CCC   .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
 4  GGA...:....:....:....:...UCC   .    .    .    .    .    .    .    .    .    .    .    .    .    .
 5  GGU...:....:....:....:........ACC   .    .    .    .    .    .    .    .    .    .    .    .    .
 6   .    :    :  (GAG)..:.............GAG..CUC   .    .    .    .    .    .    .    .    .    .    .
 7   .   GCG...:....:....:..............:....:...CGC   .    .    .    .    .    .    .    .    .    .
 8   .    :    :   GAU...:..............:....:........AUC   .    .    .    .    .    .    .    .    .
 9   .    :    :    .    :    .    .    :   CUG............CAG   .    .    .    .    .    .    .    .
10   .    :   GUU........:..............:....:..................AAC   .    .    .    .    .    .    .
11   .    :    :    .    :    .    .    :   CUU......................AAG   .    .    .    .    .    .
12   .   GCA...:.........:..............:....:............................UGC   .    .    .    .    .
13   .    .    :    .    :    .    .   GAA.....................................UUC   .    .    .    .
14   .    .   GUG........:..........................................................CAC   .    .    .
15   .    .   GUA........:...........................................................:...UAC   .    .
16   .    .    .    .    :    .    .    .    .    .    .    .    .    .    .    .   CAU.......AUG   .
17   .    .    .    .   CCA........................................................................UGG
```

Fig. 1. Amino acid and codon chronology, 20 amino acids and 17 codon pairs. Only the most stable codons are sampled, with their respective complementary codons. This makes all together 17 complementary pairs. The amino acid chronology is calculated on the basis of total 46 criteria, as in [9], after addition of two more criteria based on composition of the earliest proteins [11], and on the complementary code of Arques and Michel [22]. Two steps of filtering are applied for derivation of the consensus amino acid chronology [8].

This may serve as a 'glycine clock' for construction of rooted phylogenetic trees [11].

3. Stages of the evolution of protein structure

3.1. *Short peptides*

Let us return to Fig. 1. The next two codons after GCC and GGC are apparently derived from the first two by transitions of G to A and/or C to U in the middle positions of the GCC and GGC triplets. Thus, the middle purine codons GGC and GAC will stay in the same ('Gly') strand as well as the middle pyrimidine codons GCC and GUC would stay in the 'Ala' strand. All subsequent codons are derived via mutations in the redundant third positions of already acquired triplets, thus keeping all middle purine codons in the Gly strand, and all middle pyrimidine codons in the Ala strand. Therefore, at this early stage of molecular evolution, there would be two kinds of mixed sequence peptides of two different alphabets corresponding to aRb and aYb triplets. The later more advanced polypeptide chains are likely to emerge via fusion of respective minigenes of these two kinds. The protein sequences would then appear as a mosaic of the two alphabets with the elements of, presumably, certain optimum size.

This expectation is recently confirmed by cross- and autocorrelation analysis of 23 bacterial proteomes, which revealed the 12-residue periodicity of two alternating alphabets [9]. Thus the elementary mosaic units were six residues long. We may now outline two of the earliest stages of molecular evolution: (i) $(GGC)_6 \cdot (GCC)_6$ RNA duplexes encoding $(Gly)_6$ and $(Ala)_6$ homopeptides, and (ii) the six codon minigenes encoding peptides of Gly and Ala alphabets.

3.2. *Close loop stage*

With a gradual increase of the lengths of the evolving protein chains, the stage should have been reached when the flexible polypeptide chains would frequently make loops with the ends coming in contact. This loop (ring) closure phenomenon well known in polymer physics [12] is characterized by the optimum contour length of the loops, about 25–30 residues in the case of proteins [13]. These standard size loops were, indeed, discovered [13–16] to be a major building unit in proteins. These loops have been found, indeed, to immediately follow one after another along the sequences. This reflects the third and fourth stages in protein evolution: a loop closure stage and a loop fusion stage. The loop closure stage is also reflected in protein sequences as a 25–30-residue autocorrelation distance between hydrophobic amino acids [9,16].

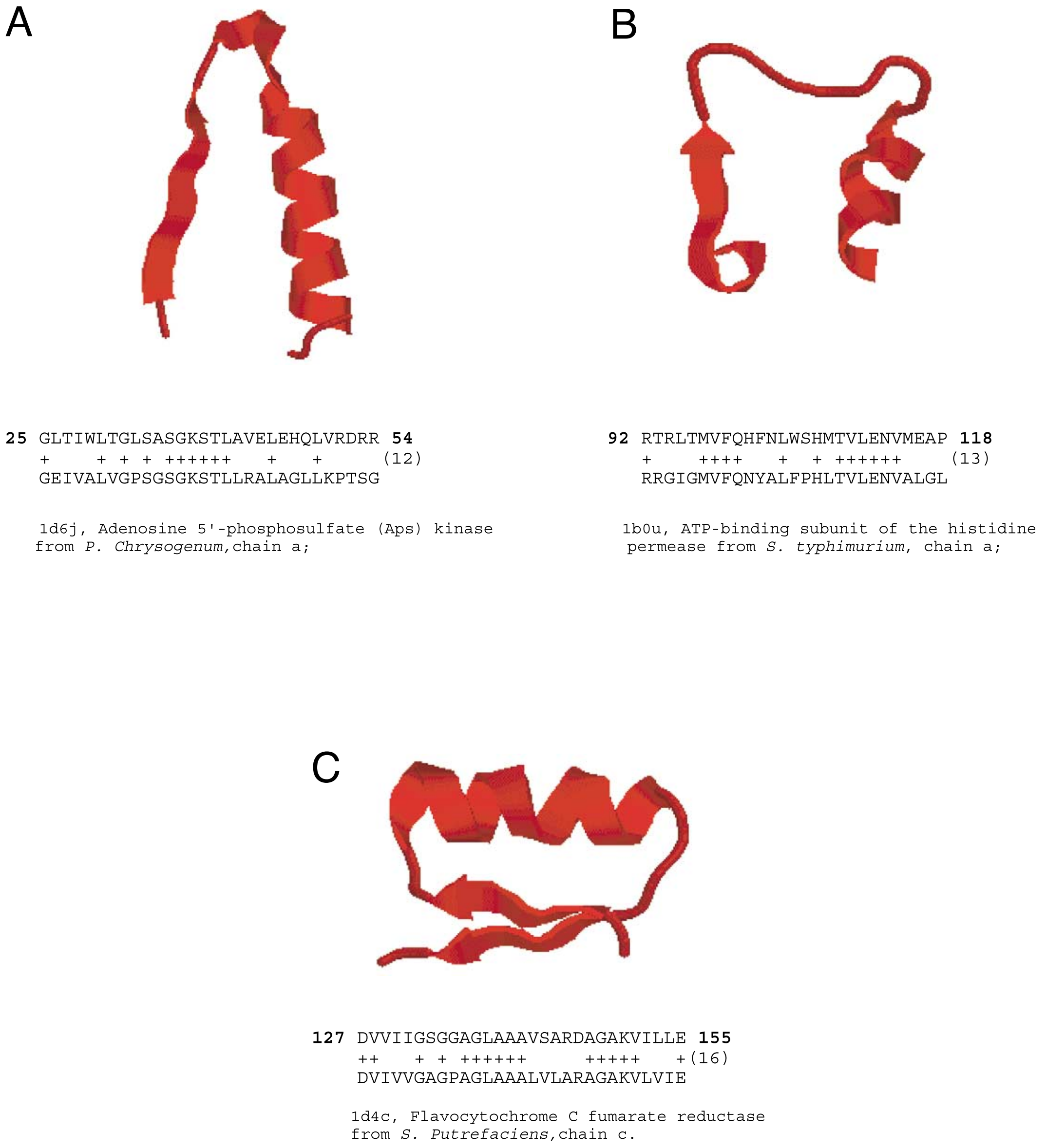
The discovery of the closed loops linearly arranged along the sequences has far-reaching implications. In particular, one could imagine that the earliest loop size proteins had specific sequences, the mutational versions of which may still be found in modern proteins. A massive search for such sequence prototypes in the complete bacterial proteomes is under way. Fig. 2 shows three examples of already detected prototypes. Remarkably, the descendants of the ancient prototype sequences are found in closed loops as anticipated.

3.3. *Fold (domain) stage and modern multidomain proteins*

How many of the closed loops would make a protein globule? Typical folds of the proteins, as well as the single-fold proteins of most frequent sizes contain 100–200 amino acid residues [17,18] that corresponds to three to five closed loops. It is not immediately clear what provides this size limit. One

Table 1
Stages of protein evolution

	Description of the stage	Characteristic size
I.	homopeptides of Ala and Gly encoded by (GCC)·(GGC) duplexes	six amino acids
II.	mixed peptides of two alphabet types	six amino acids
III.	chains of optimal length close the ends by interactions between amino acid residues [13]	25–35 amino acids
IV.	the loops are joined in linear arrays and form folds (domains)	100–200 amino acids
V.	modern multidomain proteins are formed	$(100–200)_n$ amino acids



```
25 GLTIWLTGLSASGKSTLAVELEHQLVRDRR 54
   +    + + + ++++++   +   +      (12)
   GEIVALVGPSGSGKSTLLRALAGLLKPTSG
```

1d6j, Adenosine 5'-phosphosulfate (Aps) kinase from *P. Chrysogenum*,chain a;

```
92 RTRLTMVFQHFNLWSHMTVLENVMEAP 118
   +    ++++   +  + ++++++    (13)
   RRGIGMVFQNYALFPHLTVLENVALGL
```

1b0u, ATP-binding subunit of the histidine permease from *S. typhimurium*, chain a;

```
127 DVVIIGSGGAGLAAAVSARDAGAKVILLE 155
    ++   + + ++++++     +++++   +(16)
    DVIVVGAGPAGLAAALVLARAGAKVLVIE
```

1d4c, Flavocytochrome C fumarate reductase from *S. Putrefaciens*,chain c.

Fig. 2. Examples of most frequent sequence motifs of prokaryotic proteins, and their representative structures. The motifs correspond to those sequences for which large amounts of matching sequence segments are found in 23 bacterial proteomes. In the plane projections of the loops are shown. Sequences, their positions in the respective crystallized protein, number of matches to the prototypes, and PDB descriptions of the proteins are indicated for each loop.

attractive explanation involves the ring closure again: this time of double-stranded DNA molecules. It is physically as inevitable as the protein loop closure. In both cases the loop (ring) formation provides greater stability to the molecules, which is of an obvious evolutionary advantage. The optimum DNA circularization size is about 300–600 bp [12,19,20] that corresponds to the typical protein fold size as above. The domain size analysis of crystallized proteins shows that 150 amino acid residues is a major mode in the size distributions [21].

Present-day multidomain proteins are constructed from linearly fused domains/folds of the typical size and represent the most recent stage in the evolution of the gene/protein. The five major stages of protein evolution are outlined in Table 1. Each has its own physical size scale. This scheme summarizes the theory of early molecular evolution developed on the basis of consecutively speculations, predictions and confirmations as briefly outlined above.

Acknowledgements: Thanks are due to Mrs. A. Weinberg for editing the manuscript. I.N.B. is a post-doctoral fellow of the Feinberg Graduate School, The Weizmann Institute of Science.

References

[1] Eigen, M. and Winkler-Oswatitsch, R. (1981) Naturwissenschaften 68, 217–228.
[2] Eigen, M. and Winkler-Oswatitsch, R. (1981) Naturwissenschaften 68, 282–292.
[3] Shepherd, J.C.W. (1981) Proc. Natl. Acad. Sci. USA 78, 1596–1600.
[4] Trifonov, E.N. and Bettecken, T. (1997) Gene 205, 1–6.
[5] Trifonov, E.N. (1987) J. Mol. Biol. 194, 643–652.
[6] Lagunez-Otero, J. and Trifonov, E.N. (1992) J. Biomol. Struct. Dyn. 10, 455–464.

[7] Miller, S.L. (1987) Cold Spring Harb. Symp. Quant. Biol. 52, 17–27.

[8] Trifonov, E.N. (2000) Gene 261, 139–151.

[9] Trifonov, E.N., Kirzhner, A., Kirzhner, V.M. and Berezovsky, I.N. (2001) J. Mol. Evol. 53, 394–401.

[10] Eigen, M. and Schuster, P. (1978) Naturwissenschaften 65, 341–369.

[11] Trifonov, E.N. (1999) Gene Ther. Mol. Biol. 4, 313–322.

[12] Shimada, J. and Yamakawa, H. (1984) Macromolecules 17, 689–698.

[13] Berezovsky, I.N., Grosberg, A.Y. and Trifonov, E.N. (2000) FEBS Lett. 466, 283–286.

[14] Berezovsky, I.N. and Trifonov, E.N. (2001) J. Mol. Biol. 307, 1419–1426.

[15] Berezovsky, I.N. and Trifonov, E.N. (2001) Protein Eng. 14, 403–407.

[16] Berezovsky, I.N., Kirzhner, V.M., Kirzhner, A. and Trifonov, E.N. (2001) Proteins Struct. Funct. Genet. 45, 346–350.

[17] Jones, S., Stewart, M., Michie, A., Swindells, M.B., Orengo, C. and Thornton, J.M. (1998) Protein Sci. 7, 233–242.

[18] Wheelan, S.J., Marchler-Bauer, A. and Bryant, S.H. (2000) Bioinformatics 16, 613–618.

[19] Shore, D., Langowski, J. and Baldwin, R.L. (1981) Proc. Natl. Acad. Sci. USA 78, 4833–4837.

[20] Trifonov, E.N. (1995) J. Mol. Evol. 40, 337–342.

[21] Gerstein, M. (1998) Fold Des. 3, 497–512.

[22] Arques, D.G. and Michel, C.J. (1996) J. Theor. Biol. 182, 45–58.