J Mol Evol (2001) 53:394–401
DOI: 10.1007/s002390010229

JOURNAL OF MOLECULAR EVOLUTION

Distinct Stages of Protein Evolution as Suggested by Protein Sequence Analysis

Edward N. Trifonov,[1,2] Alla Kirzhner,[2] Valery M. Kirzhner,[2] Igor N. Berezovsky[1]

[1] Department of Structural Biology, The Weizman Institute of Science, Rehovot 76100, Israel
[2] Genome Diversity Center, Institute of Evolution, University of Haifa, Haifa 31905, Israel

Received: 21 December 2000 / Accepted: 28 February 2001

Abstract. Evolution of proteins encoded in nucleotide sequences began with the advent of the triplet code. The chronological order of the appearance of amino acids on the evolution scene and the steps in the evolution of the triplet code have been recently reconstructed (Trifonov, 2000b) on the basis of 40 different ranking criteria and hypotheses. According to the consensus chronology, the pair of complementary GGC and GCC codons for the amino acids alanine and glycine appeared first. Other codons appeared as complementary pairs as well, which divided their respective amino acids into two alphabets, encoded by triplets with either central purines or central pyrimidines: G, D, S, E, N, R, K, Q, C, H, Y, and W (Glycine alphabet *G*) and A, V, P, S, L, T, I, F, and M (Alanine alphabet *A*). It is speculated that the earliest polypeptide chains were very short, presumably of uniform length, belonging to two alphabet types encoded in the two complementary strands of the earliest mRNA duplexes. After the fusion of the minigenes, a mosaic of the alphabets would form. Traces of the predicted mosaic structure have been, indeed, detected in the protein sequences of complete prokaryotic genomes in the form of weak oscillations with the period 12 residues in the form of alteration of two types of 6 residue long units. The next stage of protein evolution corresponded to the closure of the chains in the loops of the size 25–30 residues (Berezovsky et al., 2000). Autocorrelation analysis of proteins of 23 complete archaebacterial and eubacterial genomes revealed that the preferred distances between valine, alanine, glycine, leucine, and isoleucine along the sequences are in the same range of 25–30 residues, indicating that the loops are primarily closed by hydrophobic interactions between the ends of the loops. The loop closure stage is followed by the formation of typical folds of 100–200 amino acids, via end-to-end fusion of the genes encoding the loop-size chains. This size was apparently dictated by the optimal ring closure for DNA. In both cases the closure into the ring (loop) rendered evolutionarily advantageous stability to the respective structures. Further gene fusions lead to the formation of modern multidomain proteins. Recombinational gene splicing is likely to have appeared after the DNA circularization stage.

Key words: Early evolution — Amino acid chronology — Codon chronology — Triplet code — Homopeptides — Heteropeptides — Sequence mosaic — Closed loops — Autocorrelation — DNA ring closure — Protein folds — Multidomain proteins — Gene splicing

Introduction

A frequent argument in support of the idea about the improbability of life is an estimate of the likelihood of generating, by chance, a meaningful protein sequence of, for example, 100 amino acid residues. The number of possible different molecules of this size is on the order of 10^{130}, much larger than, say, the number of atoms in the visible Universe. This estimate is, however, a self-

Correspondence to: E.N. Trifonov; *present address:* Department of Structural Biology, The Weizmann Institute of Science; *email:* edward.trifonov@weizmann.ac.il

inflicted scare, since life most certainly began with much smaller molecules. For example, extant biologically active peptides quite frequently have the size of 20–40 residues (Andreu and Rivas, 1998), and functional minigenes are known to be even as small as 2 codons (Hernandez-Sanchez et al., 1998; Dincbas et al., 1999). Thus, it is quite reasonable to apply the general evolutionary formula "from simple to more complex" to protein structure and sequence as well, and to explore the possible simple precursor stages in protein evolution. Evidence of such simple stages is provided, for the first time, by the autocorrelation analyses of the prokaryotic protein sequences, which will be described. Thus, the foundation has been established for describing several distinct stages in the evolution of nucleic acid-encoded proteins, from simple to complex, reflected both in the structure of modern proteins and in their sequences. The suggested stages are: (I) short homopeptides Gly_n and Ala_n, (II) mixed-sequence 6 residue-long peptides of Glycine (*G*) and Alanine (*A*) alphabets, (III) 25–30 residue-long peptides closed into the loops by the end-to-end contacts, (IV) 100–200 amino acid residue protein folds (domains), and (V) multidomain proteins. Gene splicing presumably enters at stages IV–V "politely" (Zuckerkandl 1986), in such a way that the earlier size regularities are still respected.

Results and Discussion

Homopeptide Stage. Evolution of the Triplet Code

The range of amino acids of the earliest proteins was presumably very limited. A natural question to ask is which amino acids were the first to appear, and in what chronological order did all the other amino acids appear? Each of the numerous theories on the origin of the genetic code, and various other considerations, suggest that amino acids have a certain temporal order. Some theories support others, whereas some are contradictory. One could think of a balanced consensus that would take into account all the various estimates. However, the question about weights to be given to the criteria immediately arises, making every attempt of this kind inevitably subjective and doubtful.

One fair and reasonable way to derive the consensus is rank-averaging of different criteria, with no weights given, except for eliminating those that are almost identical, thus combining them in one criterion. Such an analysis is indeed performed with 40 different criteria of amino acid chronology (Trifonov, 2000b). Among the criteria are various theories suggesting the existence of a specific order within the amino acids and/or codons, like the coevolution theory of Wong (1981), the RNA theory of Eigen et al. (1981), Jukes' theory (1973), and other hypotheses. Criteria based on the chemical simplicity of amino acids, on their reactivity, or on the composition of early proteins, and other factors are included as well. Taking all this into consideration, a striking order in both amino acid chronology and derived codon chronology is revealed: (1) amino acids of the imitation experiments by S. Miller (1953, 1987) are in the lead positions; (2) the codons appear to have been engaged as complementary pairs; (3) More stable codon pairs are engaged first; and (4) new codons are simply point-change derivatives of the previously engaged ones.

Table 1. Consensus chronology of amino acids calculated on the basis *of 44 criteria*

	Average rank	±	Order
G	4.7	0.8	1
A	5.2	0.9	2
V	6.8	0.7	3
D	7.3	0.7	4
S	8.0	0.7	5
E	8.5	0.7	6
P	8.8	0.8	7
L	9.7	0.8	8
T	10.2	0.6	9
N	11.5	0.7	10
R	11.5	0.7	11
I	11.6	0.7	12
K	11.8	0.8	13
Q	11.9	0.7	14
C	12.3	0.8	15
F	12.3	0.8	16
H	13.1	0.7	17
M	14.3	0.6	18
Y	14.4	0.6	19
W	15.8	0.6	20

One filterinig step is used (Trifonov 2000b), reducing the number of independent criteria to 36.

An updated version of the amino acid and codon chronologies is presented in Table 1 and Fig. 1. These result from calculations performed the same way as in the above cited work, with the larger number of criteria. The added criteria of the amino acid chronology are based on the reconstructed amino acid composition of ancient ferredoxin (Eck and Dayhoff 1966), the stability of amino acids in a reducing atmosphere (Eck and Dayhoff 1966 and references therein), the mutational stability of codon repertoires (Luo, 1988), and on the molecular volumes of amino acids (Haig and Hurst, 1991). Although there is some uncertainty (see, for example, the average rank values for P and E, and for I, R, and K in Table 1), the estimates of the rank averages for most of the amino acids allow them to be ordered rather uniquely in the consensus chronology. Remarkably, the amino acids of Miller's mix (Miller 1953, 1987), G, A, V, D, S, E, P, L, and I, within the accuracy of the estimates, are all found at the top of the list.

This would mean that emerging life utilized first, of all those amino acids that were already present in the environment, which is evidence for the fundamentally opportunistic nature of the life process.

```
     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 Glu Leu Thr Asn Arg Ile Ser Arg Lys Gln Leu Cys Phe His Met Tyr Trp
                        ucx     cux         cgx     agy agr         uux                 trm trm
  1 GGC.GCC  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
  2  :   :  GUC.GAC  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
  3 GGG..:...:...:..CCC  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
  4 GGA..:...:...:...:..UCC  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
  5  :   :   :  gag..:...:..GAG.CUC  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
  6 GGU..:...:...:...:...:...:...:..ACC  .   .   .   .   .   .   .   .   .   .   .   .   .   .
  7  .   :  GUU..:...:...:...:...:...:..AAC  .   .   .   .   .   .   .   .   .   .   .   .   .
  8  .  GCG..:...:...:...:...:...:...:...:..CGC  .   .   .   .   .   .   .   .   .   .   .   .
  9  .   :   :   :  CCG..:...:...:...:...:..CGG  .   .   .   .   .   .   .   .   .   .   .   .
 10  .   :   :   :   :  UCG..:...:...:...:..CGA  .   .   .   .   .   .   .   .   .   .   .   .
 11  .   :   :   :   :   :   :   :  ACG..:..CGU  .   .   .   .   .   .   .   .   .   .   .   .
 12  .   :   :  GAU..:...:...:...:...:...:......AUC  .   .   .   .   .   .   .   .   .   .   .
 13  .  GCU..:.......:...:...:...:...:...:.......:..AGC  .   .   .   .   .   .   .   .   .   .
 14  .   :   :   .   :   :   :   :  ACU..:.......:..AGU  .   .   .   .   .   .   .   .   .   .
 15  .   :   :   .  CCU..:...:...:...:...:.......:......AGG  .   .   .   .   .   .   .   .   .
 16  .   :   :   .   :  UCU..:...:...:...:.......:......AGA  .   .   .   .   .   .   .   .   .
 17  .   :   :   .   :   :   :  CUU..:...:.......:..........AAG  .   .   .   .   .   .   .   .
 18  .   :   :   .   :   :   :  CUG..:...:.......:...........:..CAG  .   .   .   .   .   .   .
 19  .   :   :   .   :   :   :   :   :   :       :           :  CAA.UUG  .   .   .   .   .   .
 20  .  GCA..:.......:...:...:...:...:...:.......:...........:.......:. UGC  .   .   .   .   .
 21  .   .   :   .   :   :   :   :  ACA..:.......:...........:.......:..UGU  .   .   .   .   .
 22  .   .   :   .   :   :  GAA..:.......:.......:...........:.......:......UUC  .   .   .   .
 23  .   .   :   .   :   :   .   :   .   :       :   .   .  AAA......:......UUU  .   .   .   .
 24  .   .  GUG......:...:.......:.......:.......:...................:..........CAC  .   .   .
 25  .   .   :   .   :   :   .   :   .   :       :   .   .   .   .   :   .   .  CAU.AUG  .   .
 26  .   .  GUA......:...:.......:.......:.......:...................:..................UAC  .
 27  .   .   .   .   :   :   .  CUA......:.......:...................:..................uag  .
 28  .   .   .   .   :   :   .   .   .   :      AUA..................:..................UAU  .
 29  .   .   .   .   :   :   .   .   .  AAU.....AUU                  :                   :   .
 30  .   .   .   .   :   :   .   .   .   .   .   .   .   .   .   .  UUA.................uaa  .
 31  .   .   .   .  CCA..:..................................................................UGG
 32  .   .   .   .   .  UCA.................................................................uga
```

Fig. 1. Reconstruction of the codon chronology. The upper line corresponds to the consensus chronology of the amino acids. Numbered lines contain complementary codon pairs. Codons of the same repertoire (for the same amino acid) are arranged vertically.

Figure 1 presents the reconstruction of the temporal order of the codons, based on the amino acid chronology as discussed, and on the original suggestion by Eigen and Schuster (1978) on the primacy of thermostability and complementarity. The figure is a modified version of the earlier scheme (Trifonov 2000b). The order of amino acids P, S, and E is slightly changed to conform with the error bars in Table 1. The lines of the scheme correspond to complementary codon pairs. In every codon repertoire (vertical sets) the most stable codons (underlined) occupy the topmost positions. A striking feature of the scheme in Figure 1 is the below-diagonal arrangement of the thermodynamically weaker codons of the respective codon repertoires. Any substantial change in the order of the amino acids (top line) would destroy the overall triangular pattern. The triangular pattern means that the consensus amino acid chronology and the rules of thermostability and complementarity together dictate that one more most natural rule be fulfilled: new codons are derived from existing ones as single-point mutations—a processivity rule. Most of the new codons are generated by the wobble mutations and by complementary copying, whereas the pair GUC/GAC, for V and D, is derived from the transition mutation(s) in the middle positions of the earliest GGC and GCC codons. Both the wobble mutations and the transitions are the most frequent "cheapest" mutations to occur. Thus, the new codons not only evolve from the previous ones, but in addition, they do so in the most inexpensive way. The unexpected strict order revealed in the organization of the codon chronology gives a degree of confidence that the derived temporal orders largely reflect previously undisclosed steps of evolution of the triplet code.

The first complementary codon pair, GGC*GCC, should correspond to the earliest coding RNA duplex with the complementary strands GGCGGCGGC. . .GGC and GCCGCCGCC. . .GCC of uncertain length coding for GlyGlyGly. . .Gly and AlaAlaAla. . .Ala, respectively. The original amino acid composition of the "proteins" is 50% of Gly and 50% of Ala. At the later stages of the protein's evolution the original abundance of Gly and Ala is gradually reduced by other amino acids, so that the proportions of Gly and Ala eventually go down to their current 6%–8% values. The exceptional evolutionary role of Gly is illustrated by the following important observation. The decline in the Gly content is indeed observed when functionally related prokaryotic and eukaryotic protein sequences are aligned and the composition of the shared parts of the sequences is calculated (Trifonov 1999b). The proportion of the Gly in the shared residues is as high as 14%. That value of the Gly content is likely to correspond to the moment of separation between eukaryotes and prokaryotes, about 3.5 billion years (Doolittle 1997). Thus, the Gly content may serve as a clock to construct rooted evolutionary trees. A first attempt of this kind, the rooted tree for 6 major kingdoms has been recently constructed (Trifonov 1999b). It completely matches consensus chronology of the evolutionary bifurcations obtained by traditional non-rooted techniques (Doolittle 1997).

Earliest Mosaic Stage

With advances in the evolution of the code, new codons and amino acids were accommodated, so that the earliest homopeptides became heteropeptides, encoded as well by the mRNA duplexes. If the wobble and transition mutations were dominant a sufficiently long time, the newly formed codons of the strand initially encoding only glycines and, one step later, aspartic acid, would almost all contain purines G or A in the middle. These codons of the structure xRz correspond to the amino acids of the Glycine alphabet *G:* G, D, S, E, N, R, K, Q, C, H, Y and W (in chronological order). Similarly, the complementary strands with triplets xYz encode peptides of the Alanine alphabet *A:* A, V, P, S, L, T, I, F, and M. The alphabets have only one amino acid in common—serine, since it is encoded by the triplets of both types. Notably, the *G* alphabet consists largely of polar and charged residues, whereas the *A* alphabet consists of primarily hydrophobic residues.

The presumably short initial minigenes were likely to eventually fuse end-to-end, forming a mosaic of the two alphabets in the oligopeptide translation products. We speculate that the mosaic elements were a certain uniform length as were the respective minigenes (Trifonov 2000). In this case we could probably attempt to detect the traces of the mosaic of two alphabets, perhaps still surviving in the modern sequences. This prospect sounds like a fantasy, considering 3.9 billion years of point mutations, deletions, and insertions in the incessantly evolving protein sequences. There is, however, some hope based on two factors. First, many of the primarily hydrophobic or hydrophilic mosaic elements selected originally for their respective functions requiring hydrophobicity/hydrophilicity would perhaps survive. Second, the informational sequences, in general, and protein sequences, in particular, carry many different overlapping codes (Trifonov 1989, 1996). This provides a certain degree of sequence conservation. Ancient sequence patterns not functional anymore may remain as a part of later functional patterns, involving the same letters in the same positions. Still, the speculated mosaic signal, if it exists at all, is expected to be very weak, so that it may be detected only in very large sequence ensembles. This is why we have chosen for the following autocorrelation sequence analysis completely sequenced archaebacterial and eubacterial genomes, 23 in all, containing about 50,000 protein sequences, including translated unidentified reading frames.

We assumed, not without a reason, that the speculated ancient mosaic would be largely an alternating type, forming a period of two mosaic units of the different alphabets. The preferred distance between letters of two different alphabets would then be equal to one unit, three units, and so on, probably decaying with the distance because of the destructive influence of deletions and insertions. Similarly, the preferred distances between letters of the same alphabet would be equal to two units, four units, etc.

What would be the size of the mosaic unit? In other words, what was the size of the very first homopeptides—oligoalanines and oligoglycines? The upper limit to that size is set by the solubility of the peptides, which is 7–8 residues (Ogata, in press). The lower limit can be derived from the RNA duplex stability. The G+C rich RNA duplex of a length as short as 10 base pairs is stable at 90–100 C° (Frank-Kamenetskii, 1990). This corresponds to the peptide size of the 3–4 residues. Thus, the expected size of the speculated mosaic element would be in the range of 3–8 amino acid residues.

To our delight, the predictions about the hidden periodicity present in the protein sequences was fully confirmed by analyzing the proteins of prokaryotic complete genomes. Fig. 2 presents the autocorrelation functions showing the preferred distances between the letters of alphabets *G* and *A* (common S is excluded) in the proteins. The decaying oscillations are clearly seen in the smoothed curves. The smoothing is necessary to eliminate oscillation due to a 3.5 residue, short-range periodicity of the α-helical regions (e. g., Herzel et al. 1999).

Alternating minima and maxima are seen, as predicted, at distances 6, 12, and about 19 residues. This corresponds to the size of a detected mosaic unit of about 6.3 residues, well within the estimated range of 3 to 8 residues. It does not have to be an integer, if unequal rates of the insertions and deletions are assumed, which would make the apparent unit size proportionally larger or smaller. One could also imagine that the original minigenes had a range of sizes, within the estimated 3 to 8 residues for the peptides, with the non-integer average close to 6.

Loop Closure Stage

A fundamental unit of the protein structure—closed loops of a typical size of 25–30 residues—has been recently discovered (Berezovsky et al. 2000). Contrary to the traditional approach to protein structure based on the primacy of secondary structure elements, e.g., α-helices, β-sheets, and turns, in this work the protein globule is viewed as a path of the polypeptide chain trajectory, with a hierarchy of returns (closed loops) and chain-to-chain contacts. The contour length distribution of the closed loops of the globular proteins shows one major maximum, at 25–30 amino acid residues. Analysis of repre-

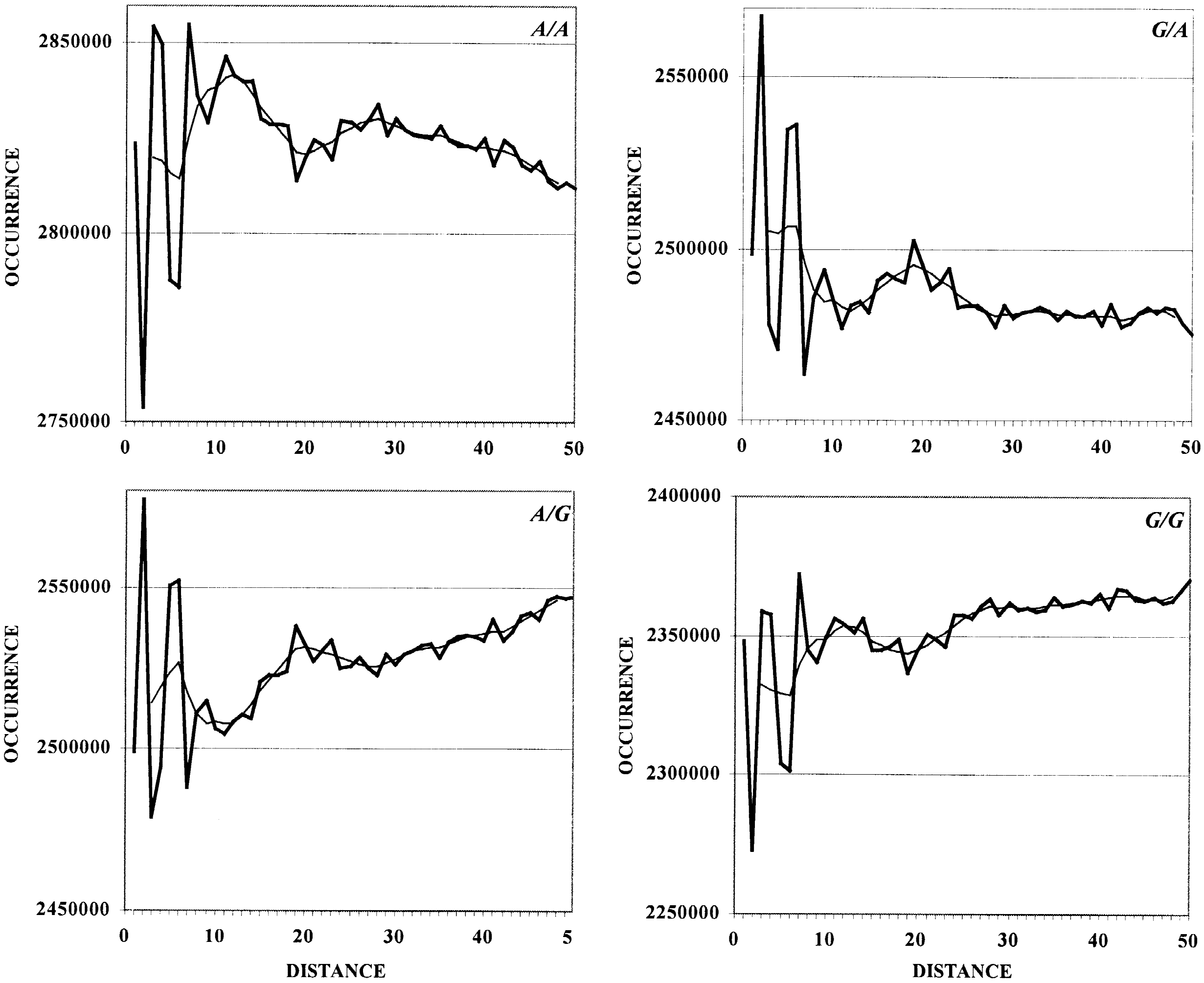

Fig. 2. Traces of the two-alphabet mosaic structure in modern protein sequences. Positional correlations and autocorrelations between residues of alphabet *G* (G, D, E, N, R, K, Q, C, H, Y, and W) and alphabet *A* (A, V, P, L, T, I, F, M). Smoothed curves show extrema at 6, 12, and 19 residues. Protein sequences from the 23 complete prokaryotic genomes were taken for the calculations.

sentatives of 10 major fold types also reveals that the nearly standard closed loops follow one after another along the polypeptide chain making linear arrays of the loops (Berezovsky et al. 2000). Such an arrangement could have resulted from end-to-end fusions of small genes encoding the standard loop-size polypeptides. All intraglobular residue-to-residue contacts close the respective loops—segments of the chain between the loop ends. The stability of the globule is provided largely by these contacts, which consist of van der Waals locks between the meeting loop ends. The primary loops of the standard size were found to be mostly closed by hydrophobic residues (Berezovsky and Trifonov 2001a).

What is the origin of the 25–30-residue unit-size loops? The plausible answer comes from polymer statistics. The polypeptide chains possess some flexibility due to significant rotational freedom around the bonds of the backbone. Because of this flexibility, the ends of sufficiently long chains may come together and make the closed loop. In the partially structured loop with 40%–50% of the residues involved in the rigid α-helices as in the natural proteins, an optimal loop closure size of 20–35 residues is statistically predicted, as was indeed observed (Berezovsky et al. 2000 and references therein). Actually, the loop closure is an inevitable stage for the evolving protein chains to go through. The must of the loop closure also carries with it an important selective advantage: higher stability of the polypeptide chains closed in the loops.

Crucial evidence that the closed loops indeed have been accommodated in the evolutionary scenario could be taken from the protein sequences. In particular, since some residues, presumably the hydrophobic ones, would be the better choice for the loop-closing interactions, they would be selected for this important function. In this case the preferred loop sizes may also appear as the preferred distances between these residues along the protein sequences. Such a sequence pattern would again reflect the very distant past of the protein's evolution. One would need a very large ensemble of the sequences in order to detect this possible pattern. For the analysis we have taken protein sequences of 23 complete pro-

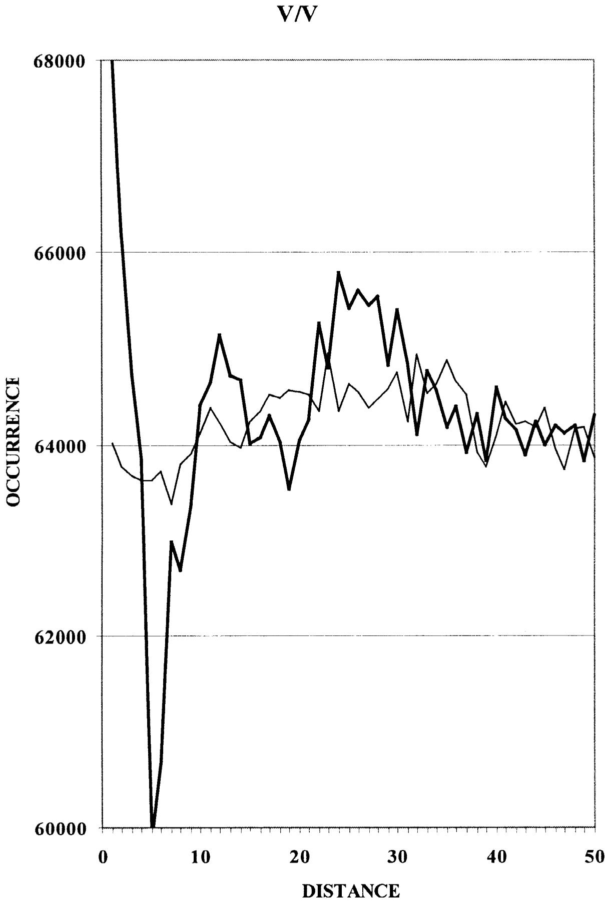

Fig. 3. Positional autocorrelation for valines in protein sequences of 23 prokaryotic genomes. Thick line—natural sequences, thin line—shuffled sequences. Note the maximum at 22–31 residues.

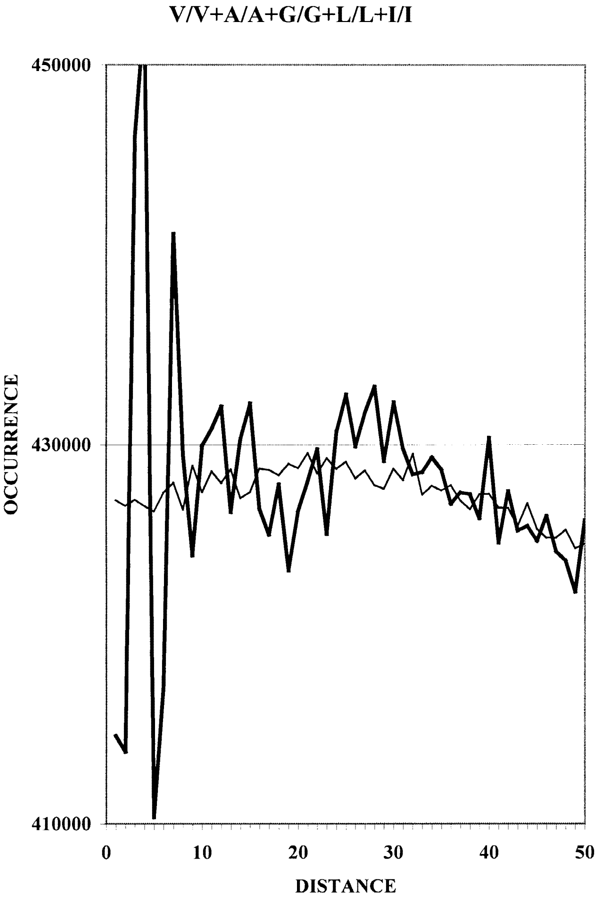

Fig. 4. Sum of positional autocorrelations for residues V, A, G, L, and I in protein sequences of 23 prokaryotic genomes. Thick line—natural sequences, thin line—shuffled sequences. Note the maximum at 25–30 residues.

karyotic genomes, available in the beginning of this study. The sequence distances (positional autocorrelation functions) have been calculated individually for all 20 amino acids. To our satisfaction, many of them displayed the expected preference to the residue-to-residue distance within the range of 25–30 amino acids.

Fig. 3 shows the distance distribution for valine, which appeared to be the strongest residue in this respect. The curve has a distinct maximum in the range of 22–31 residues, with the amplitude well beyond random fluctuations (compared with the curve calculated for shuffled sequences). Alanine, glycine, leucine, and isoleucine are the next largest contributors to the effect. Fig. 4 shows the combined curve for all five residues, further supporting the conclusion that the protein sequence manifests the preferred distance of 25–30 residues for some amino acids, as expected. The fact that the major contributors are all hydrophobic provides an additional strong support to the view that the proteins, in their evolution, passed through the loop closure stage, reflected both in the typical closed loop size and in the distribution of hydrophobic residues, presumably locking the loop ends. Of course, other residues may also be involved in the locks, which are generally stabilized by van der Waals interactions (Berezovsky and Trifonov 2001b).

Formation of Domains (Folds), Multidomain Proteins

Fusion of the genes encoding the loop-size polypeptides results in the formation of larger proteins with a developed hierarchy of domains and subdomains (Berezovsky et al. 1999), due to a variety of interactions between the primary loops. The longer genes would eventually reach the size when DNA, in its turn, will have to close into the ring, by fusion of its ends. The inevitability of this event is dictated by the laws of polymer statistics and stability advantage, which is also true with the polypeptide loop closure. Since the DNA molecule is more rigid, the optimal length for the DNA circularization is substantially larger, about 300–600 base pairs (Shore et al. 1981) corresponding size-wise to about 100–200 encoded amino acid residues, or to 4–7 standard size closed loops. This range of sizes is characteristic of typical domain representatives—protein folds (Wheelan et al. 2000). It is likely that this size is established at the gene circularization stage. A more detailed analysis of the sizes of the protein chains (Berman et al. 1994) results in somewhat different values for the protein chain lengths in eukaryotes—about 120 residues, and in prokaryotes about 150 residues. Both correspond to the optimal range for the DNA ring closure. The epoch of DNA circularization apparently had its impact not only on the protein sizes, but also on the genome structure in general. For example, mobile DNA elements, not necessarily protein-coding, prefer the ring closure sizes and multiples thereof (Trifonov 1997). One can speak in general about genome units of the optimal DNA circularization size from which modern genomes are built, by recombinational fusion of the initially separate DNA rings (Trifonov 1995). Similar estimates of the genome unit sizes, consistent with the above values, are also obtained from the distribution of methionines in the protein sequences, from the distribution of rare triplets along mRNA and other data (reviewed in Trifonov 1999a).

For all earlier stages of protein evolution as described here, the forthcoming stage always involves end-to-end fusion of the protein-coding genes of a previous stage. Likewise, in the last, close-to-current stage, respective genome unit-size genes have fused to form large genes coding for multidomain proteins with several typical fold-size domains, sometimes forming rather long linear arrays.

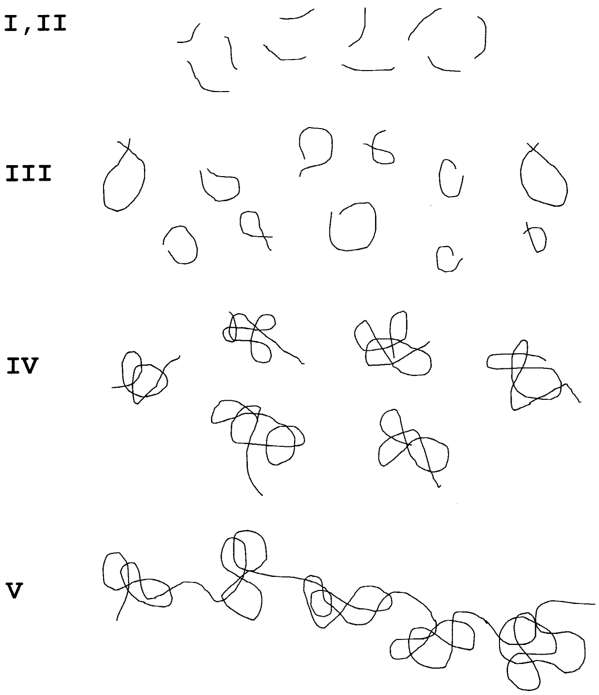

Fig. 5. A scheme of five major stages of protein evolution. I, II—six residues long homo- and heteropeptides. III—loop closure stage. IV—protein folds. V—multidomain proteins.

Gene Splicing

Undoubtedly, gene splicing is an important evolutionary invention that presumably allows for reshuffling of the coding sequences and thereby increases the frequency of the recombination events (Gilbert 1978). In addition, it allows for spatial separation of the otherwise conflicting sequence messages responsible for the chromatin structure and for the protein coding (Zuckerkandl 1981; Trifonov 1993; Denisov et al. 1997). It is thought that due to this separation, the quality (performance) of the encoded protein is improved. To be consistent with the characteristic sizes pertinent to the sequential stages of protein evolution, the newly inserted coding sequences have to obey the size distribution rules. In particular, the typical sizes of the closed loops and the folds should be maintained. That is, there should be a positional correlation between the exon-exon junctions in the protein sequences on one hand, and the borders between the closed loops and domain-domain borders on the other. Such a correlation is observed when we consider the so-called centripetal modules of haemoglobin (Go 1981), enolase (Roy et al. 1999), and xylanase (Sato et al. 1999). The modules mapped on the sequence are frequently separated by the exon-exon junctions. Interestingly, the average size of the centripetal module of enolase is about 27 residues, in good agreement with the typical size of the closed loops. The most frequent size of exons, 35–37 residues (Long et al. 1995) is also rather close to the loop sizes. Thus, it seems plausible that the gene splicing does respect the loop organization of the proteins. A detailed comparison of the sequence locations of the closed loops with the positions of exon-exon junctions would be necessary to confirm the “politeness” of the introns.

Conclusion

The stages of protein evolution as outlined here are schematically illustrated in Fig. 5. Since this is the first scheme of this kind, it may be modified in future studies by acquiring new details and corrections. Importantly, it reflects the current state of our knowledge about the early evolution of the protein and may serve as a working hypothesis for further efforts in elucidating the evolution of life. The scheme is based on the initial notion that the evolution of proteins underwent distinct structural stages, from simple to complex, from short to long, having a certain characteristic size at every stage. In this respect the gene splicing does not represent a separate stage. It was perhaps invented by nature at the stage of the fold formation, or later, and apparently did not introduce any substantial change in the size-based scheme in Fig. 5.

Sequences and Calculation Methods

The protein sequences of the following complete prokaryotic genomes were used for the calculations. Archaea: *A. pernix, A. fulgidus, M. thermoautotrophicum,* and *P. horikoshii.* Eubacteria: *A. aeolicus, B. burgdorfe-*

rii, C. jejuni, C. muridarum, D. radiodurans, E. coli, H. influenzae, H. pylori, M. tuberculosis, M. pneumoniae, N. meningitidis, R. prowazekii, Synechocystis, T. maritima, T. pallidum, U. urealyticum, V. cholerae, and *X. fastidiosa.* The sequences were provided by the National Center for Biotechnology Information, via Entrez Browser. The sequences were used without any filtering, as indicated.

The positional cross- and autocorrelation functions—histograms of all encountered distances between the specified amino acids—were calculated up to a distance of 50 residues. The last 50 residues of every sequence were not taken as starting points, to avoid end-effects. The functions were calculated first for individual genomes and then summed together in one plot.

Smoothing of the curves in Figure 2 was performed by two cycles of averaging by running a window of 3 residues.

The shuffling of the sequences for all 23 genomes was performed by replacing every second randomly chosen residue by another residue within a sliding interval of 10 amino acids.

Acknowledgments. The authors are grateful to Prof. E. Nevo for encouragement and help. I.N.B. is a postdoctoral fellow of the Feinberg Graduate School at the Weizmann Institute of Science. V.M.K. is supported by the Ministry of Absorption. The numerous discussions with Prof. N. Lahav are highly appreciated.

References

Andreu D, Rivas L (1998) Animal antimicrobial peptides: an overview. Biopolymers 47:415–419

Berezovsky IN, Namiot VA, Tumanyan VG, Esipova NG (1999) Hierarchy of the interaction energy distribution in the spatial structure of globular proteins and the problem of domain definition. J Biomol Struct Dyn 17:133–155

Berezovsky IN, Grosberg AY, Trifonov EN (2000) Closed loops of nearly standard size: common basic element of protein structure. FEBS Lett 466:283–286

Berezovsky IN, Trifonov EN (2001a) Evolutionary aspects of protein structure and folding. Mol Biol 35:233–239

Berezovsky IN, Trifonov EN (2001b) Van der Waals locks: loop-n-lock structure of globular proteins. J Mol Biol 307:1419–1426

Berman AL, Kolker E, Trifonov EN (1994) Underlying order in protein sequence organization. Proc Natl Acad Sci USA 91:4044–4047

Denisov DA, Shpigelman ES, Trifonov EN (1997) Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. Gene 205:145–149

Dincbas V, Heurgue-Hamard V, Buckingham RH, Karimi R, Ehrenberg M (1999) Shutdown in protein synthesis due to the expression of mini-genes in bacteria. J Mol Biol 291:745–749

Doolittle WF (1997) Fun with genealogy. Proc Natl Acad Sci USA 94:12751–12753

Eck RV, Dayhoff Mo (1966) Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science 152:363–366

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

Eigen M, Gardiner W, Schuster P, Winkler-Oswatitsch R, (1981) The origin of genetic information. Sci Am 244:88–118

Frank-Kamenetskii MD (1990) Energetics of DNA and RNA double helices. In Landolt-Bornstein series in Biophysics, vol. 1, subvol. c. Springer-Verlag, Heidelberg, pp 228–240

Gilbert W (1978) Why genes in pieces? Nature 271:501–501

Go M (1981) Correlation of DNA exonic regions with protein structural units in haemoglobin. Nature 291:90–92

Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code. J Molec Evol 33:412–417

Hernandez-Sanches J, Valades JG, Herrera JV, Ontiveros C, Guarneros G (1998) λ *bar* minigene-mediated inhibition of protein synthesis involves accumulation of peptidyl-tRNA and starvation for tRNA. EMBO J 17:3758–3765

Herzel H, Weiss O, Trifonov EN (1999) 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics 15:187–193

Jukes TH (1973) Possibilities for the evolution of the genetic code from a preceding form. Nature 246:22–26

Long M, Rosenberg C, Gilbert W (1995) Intron phase correlations and the evolution of intron/exon structure of genes. Proc Natl Acad Sci USA 92:12495–12499

Luo LF (1988) The degeneracy rule of genetic code. Origin of Life Evol. Biosphere 18:65–70

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 Spr Harb Symp Quant Biol 52:17–27

Ogata Y, Imai E, Honda H, Hatori K, Matsuno K (2001) Hydrothermal circulation of sea water through hot vents and contribution of interface chemistry to prebiotic synthesis. Orig Life Evol Biosph (in press)

Roy SW, Nosaka M, de Souza SJ, Gilbert W (1999) Centripetal modules and ancient introns. Gene 238:85–91

Sato Y, Niimura Y, Yura K, Go M (1999) Module-intron correlation and intron sliding in family F/10 xylanase genes. Gene 238:93–101

Shore D, Langowski J, Baldwin RL (1981) DNA flexibility studied by covalent closure of short fragments into circles. Proc Natl Acad Sci USA 78:4833–4837

Trifonov EN (1989) The multiple codes of nucleotide sequences. Bull Math Biol 51:417–432

Trifonov EN (1993) Gene splicing: spatial separation of overlapping messages. Comput Chem 17:27–31

Trifonov EN (1995) Segmented structure of protein sequences and early evolution of genome by combinatorial fusion of DNA elements. J Mol Evol 40:337–342

Trifonov EN (1996) Interfering contexts of regulatory sequence elements. CABIOS 12:423–429

Trifonov EN (1997) Segmented structure of mobile and separate DNA and RNA elements as suggested by their size distributions. J Biomol Struct Dyn 14:449–457

Trifonov EN (1999a) Elucidating sequence codes: three codes for evolution. Annal NY Acad Sci 870:330–338

Trifonov EN (1999b) Glycine clock: Eubacteria first, Archaea next, Protoctista, Fungi, Planta and Animalia at last. Gene Ther Mol Biol 4:313–322

Trifonov EN (2000a) Leap into life's beginnings. Tracking the chronology of amino acids. Sci Spectra 20:62–71

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

Wheelan SJ, Marchler-Bauer A, Bryant SH (2000) Domain size distribution can predict domain boundaries. Bioinformatics 16:613–618

Wong JT-F (1981) Coevolution of genetic code and amino acid biosynthesis. Trends Bioch Sc 6:33–36

Zuckerkandl E (1981) A general function of noncoding polynucleotide sequences. Mol Biol Rep 7:149–158

Zuckerkandl E (1986) Polite DNA: functional density and functional compatibility in genomes. J Mol Evol 24:12–27