**HOW BASICS OF PROTEIN EVOLUTION COULD HELP THE GENE FINDING.**

TRIFONOV E.N.

Department of Structural Biology, The Weizmann Institute of Science, Rehovot 76100, Israel;
e-mail: bptrifo1@weizmann.weizmann.ac.il

Keywords: gene finding, protein sequence, alignment, amino acid composition, genetic code, codons

**1. Introduction**

One elegant way to improve performance of gene-finding algorithms is to filter the predictions by aligning them with already known protein sequences [1]. This procedure requires massive pair-wise sequence alignments and is confined to cases with short indels only. That is, it is applicable to single- or two-three-domain proteins with no long indels between the domains. In reality large proportion of protein sequences is characterized by rather long indels when compared to their homologues. The long indels reside, apparently, in the interdomain linker regions [2]. The filtering by alignment is expected to substantially improve if instead of looking for matching proteins of the same multidomain lengths, the unit size single-domain sequences are used as targets for the alignments. Additional improvement could, perhaps, be achieved if during the alignments higher weights were given to conserved ancient amino-acids, in particular, glycine and proline - most frequent common residues in prokaryotic sequences aligned to eukaryotic sequences, and key elements of turns in the protein structure.

**2. 125-155 aa units of protein structure**

The standard size units of protein sequence and structure is a valuable poste restante reality that is regrettably neglected. They are seen as preferential sizes of protein sequences and multiples thereof - 120-125 aa in eukaryotes and 150-155 aa in prokaryotes [3]. The same range of sizes, 100-150 aa is the most typical for structural protein domains, typical folds [4] that is known, actually, since 1972 [5]. Finally, the translational pausing, apparently, to allow for consecutive independent folding of the protein domains occurs after first 155 codons are translated, in prokaryotes [6,7], and after about 125 codons in eukaryotes (work in progress). Typical proteins have a beaded structure, with the beads (folds, domains) of the size 125-155 residues connected in a chain [4]. One good example is beta-galactosidase that consists of five such units [8]. It is speculated that the beads, originally separate molecules, have been fused together recombinationally at some stage of evolution. This is supported by the observation that methionines, initiation residues, more frequently appear at the fusion points between the units [9]. The number of different types of the sequences forming the beads is rather limited. The estimates can be taken from the studies in which protein sequence fragments of fixed lengths were compared, basically, all-to-all, for large protein sequence databases, with a striking conclusion: the sequence fragments cluster to only 100-350 different types [10,11]. The choice of the fragment sizes in these studies, 50 aa [10] and 209 aa [11] has some justification. However, since both sizes lead to the same fundamental observation - very limited number of sequence types, more appropriate choice would be of the common order about 100 residues, perhaps 125-155 residues - the above unit size. Having classified all protein sequence regions in only few hundred types, one could use the standard sequence types for detection of the short-indel functional units (domains) in the combinatorially spliced open reading frames. The long-indel interdomain linker regions can be verified by respective two-domain analogues, single protein molecules or the same size regions extracted from longer sequences.

**3. Respect to more frequently conserved residues**

In the alignment of protein sequences some mismatches are given lower penalties, by using Dayhoff-type matrices. Matches, however, are normally given the same winning weight, irrespective of the type of the amino acid. One weighting scheme is suggested by the recent reconstruction of 10 earliest codons and 7 respective amino acids: ala, asp, gly, pro, ser, thr and val [12]. The reconstruction was based on natural expandability of (GCT)n sequences, and on universal (GCU)n pattern hidden in mRNA sequences [13]. This suggested that the very first triplets were GCU and it's 9 point change derivatives. The reconstruction of the above list of the earliest amino acids was based on the experiments of S. L. Miller, and on chemical simplicity of the amino acids. Inspection of the table of the triplet code revealed practically one-to-one correspondence between these residues and the GCU-derived codons [12]. This gives reason to believe that the earliest proteins, perhaps, long time before the separation of eukaryotes from prokaryotes, had been built from the above 7 ancient residues. At later stages, with appearance of other amino acids the domination of the seven, surely, was compromised. However, one could expect that even at the stage of separation eukaryotes-prokaryotes some of the ancient residues still prevailed. This can be checked by alignment of prokaryotic and eukaryotic sequences and comparing aa composition of the common parts (points) to the composition of eukaryotic and prokaryotic proteins. We have performed this analysis on 83 arbitrarily chosen aligned sequence pairs, scoring total 3900 matching residues. The results (in %) are presented in Table 1, where the composition values for prokaryotic and eukaryotic proteins are taken from ref. [14]. The percentages presented in the Table show that in the common (about 3 billion years old) eukaryotic-prokaryotic material the residues gly, pro, his, asp and phe (labeled by "+") are more frequent than in modern proteins. Three of them belong to the earliest alphabet. That is, the earliest amino acids still dominated at that time. Glycine and proline are remarkably dominant (150% and 30% excess, respectively). Their well known common role in the protein structure is formation of turns. The observed unusual conservation of glycines and prolines, thus, indicates that the turns are, perhaps, more important in maintaining conserved protein structure than alpha-helices and beta-sheets.

Table 1. Amino-acid composition of conserved points in eukaryotic-prokaryotic alignments.

 Prokaryotes Eukaryotes Common

 total total conserved

Ala 9.4 7.1 8.4

Arg 5.4 5.0 5.1

Asn 4.2 4.5 3.2

Asp 5.6 5.2 6.8+

Cys 1.0 2.0 0.9

Gln 3.9 4.2 2.0-

Glu 6.3 6.5 5.9

Gly 7.7 6.9 18.0+

His 2.1 2.3 2.8+

Ile 5.9 5.4 4.5

Leu 9.5 9.2 8.4

Lys 5.2 6.2 3.8-

Met 2.4 2.3 1.7

Phe 3.8 4.1 4.4+

Pro 4.4 5.2 6.3+

Ser 6.0 7.6 3.9-

Thr 5.7 5.6 4.5

Trp 1.3 1.2 1.3

Tyr 3.1 3.2 2.8

Val 7.1 6.3 5.8

Accordingly, gly/gly and pro/pro matches in the protein sequence alignments should be given higher weights. The underrepresented residues, especially gln, lys and ser (labeled by "-" in Table 1), should be given, respectively, lower weights. This would improve scores for correctly aligned sequences and make the gene finding algorithms more accurate.

**References**

1. M. S. Gelfand, A. A. Mironov and P. A. Pevzner, "Gene recognition via spliced sequence alignments" Proc. Natl. Acad. Sci. USA **93**, 9061 (1996).
2. S. Pascarella and P. Argos, "Analysis of insertions / deletions in protein structures" J. Mol. Biol. **224**, 461 (1992).
3. A. L. Berman, E. Kolker and E. N. Trifonov, "Underlying order in protein sequence organization" Proc. Natl. Acad. Sci. USA **91**, 4044 (1994).
4. E. Kolker and E. N. Trifonov, "Segments, folds and overall protein structure" In Biological Structure and Dynamics, R. H. Sarma and M. H. Sarma, Eds., p. 257 (Adenine Press, 1996).
5. B. W. Matthews, J. N. Jansonius, P. M. Colman, B. P. Schoenborn and D. Dupourque, "Three-dimensional structure of thermolysin" Nature New Biol. **238**, 37 (1972).
6. C. Makhoul and E. N. Trifonov, "Periodical recurrence of translation pause sites in mRNA and standard sizes of protein sequence segments and independently folding domains" J. Biomol. Struct. Dynam. **14**, 787 (1997).
7. E. N. Trifonov, D. A. Denisov and C. Makhoul, "Interacting sequence patterns" Math. Modelling and Sc. Comp., in press.
8. R. H. Jacobson, X.-J. Zhang, R. F. DuBose and B. W. Matthews, "Three-dimensional structure of beta-galactosidase from E. coli" Nature **369**, 761 (1994).
9. E. Kolker and E. N. Trifonov, "Periodic recurrence of methionines: fossil of gene fusion?" Proc. Natl Acad. Sci. USA **92**, 557 (1995).
10. M. Linial, N. Linial, N. Tishby and G. Yona, "Global self-organization of all known protein sequences reveals inherent biological signatures" J. Mol. Biol. **268**, 539 (1997).
11. M. Riley and B. Labedan, "Protein evolution viewed through Escherichia coli protein sequences: introducing the notion of a structural segment of homology, the module" J. Mol. Biol. **268**, 857 (1997).
12. E. N. Trifonov and T. Bettecken, "Sequence fossils, triplet expansion, and reconstruction of earliest codons" Gene **205**, 1 (1997).
13. J. Lagunez-Otero and E. N. Trifonov, "mRNA periodical infrastructure complementary to the proof-reading site in the ribosome" J. Biomol. Struct. Dynam.**10**, 455 (1992).
14. D. G. Arques and C. J. Michel, "A complementary circular code in the protein coding genes" J. Theor. Biol. **182**, 45 (1996).