pp. 1–12 (2014)

The method for identification of hierarchical organization of protein sequences

A. A. Anashkina[*] and A. N. Nekrasov[†]

Abstract — The paper describes a new method of Information Structure Analysis (ISAN) for protein sequences. The method uses a new approximation for description of protein sequences, which allows us to detect hierarchically organized elements (Information Structure Elements (ISEL)) in them. It is shown that elements of the higher hierarchy level of information structure correspond to stable elements in the spatial structure of globular proteins. Based on the ISAN method, we propose an approach to design of the new sequences of recombinant proteins. The proposed approach was confirmed experimentally in the case of obtaining functionally active recombinant proteins.

Keywords: Protein sequence, information structure, protein hierarchical organization, ISAN method, informational unit.

The idea that proteins can have a hierarchical organization appeared prior to revealing the first spatial structure of protein. Studying peptide bond, Linus Pauling had found that it has a series of steric restrictions and can form periodic structures stabilized by hydrogen bonds [19]. Thus, Pauling implicitly proposed the following scheme of hierarchical organization of protein molecules:

$$\begin{array}{c}\text{Aminoacid residue}\\ \downarrow\\ \text{Secondary structure}\\ \downarrow\\ \text{Spatial structure of protein}\end{array}$$

The idea of the hierarchy in the organization of protein molecules has found its place in the analysis of their spatial structures. A series of general regularities in the organization of globular proteins were indicated as early as in [23]. Those regularities formed the base of classification of spatial structures on the basis of mutual

[*] Engelhardt Institute of Molecular Biology of Russian Academy of Sciences, Moscow 119991, Russia. E-mail: nastya@eimb.ru

[†] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry of Russian Academy of Sciences, Moscow 117997, Russia. E-mail: alexei_nekrasov@mail.ru

The work was partly supported by the Program of Fundamental Research of the Presidium of the RAS within the strategic directions of science development for 2014 ‘Fundamental problems of mathematical modelling’ and the project ‘The mathematical model of spatial organization of natural polypeptide chains on the base of the information content of primary structure’, and by the scientific program of the Presidium of the RAS ‘Molecular and cell biology’.

arrangements of secondary structure elements, ‘folds’, in them [13, 18]. The works of A. V. Efimov also lie in this direction, he considered ‘folds’ of globular proteins as the hierarchical system, the transition between its levels is caused by addition of a new element of the secondary structure to the existing spatial organization [5]. However, further studies had shown that the same ‘folds’ (spatial packing) can be formed by completely different amino acid sequences. This fact eliminates the practical value of the work on classification of ‘folds’ and attempts to use these classifications to build proteins *de novo* had only limited success. The work focused on the analysis of local spatial organization of proteins allowed one to propose different variants of ‘structural alphabets’ of polypeptide chains with different lengths of elementary units [3, 12]. It is necessary to note that intermediate levels of structural organization between ‘folds’ and ‘letters of structure alphabets’ were not considered in those papers and in the series of similar researches. The study of regularities of protein sequences allowed one to reveal conservative combinations of amino acid residues (see, e.g., the ProSite base [6]). Note that detected conservative elements in amino acid sequences of proteins have no hierarchical organization.

However, the hierarchy of structural organization is an absolutely necessary feature of proteins. First of all, it is necessary for the folding of a polypeptide chain [2, 4, 8, 9]. Moreover, it is impossible to form a system of interactions within a globula that could provide a ‘deterministic mobility’ necessary for functioning of enzymes. It was indicated in [1] that ‘there is no doubt that all structural and functional protein characteristics are determined by their amino acid sequence’. Thus, the hierarchy in the structural organization of proteins must find its reflection in protein sequences. However, no trace of the hierarchical organization in protein sequence was found yet [20, 21, 26, 27]. And the level of edit of protein sequences was estimated as 1.

In this research we have tried to develop a method allowing one to detect a hierarchical organization in the amino acid protein sequence. Assuming that the unit in the amino acid sequence is not individual amino acid residues (a.a.r.), but a group of a.a.r. standing nearly, we have succeeded in solving this problem.

1. Materials and methods

1.1. Theoretical justification of the method

The search for the size of the unit of protein sequences was performed with the study of information organization in protein sequences. Initial data were taken from NRDB databases of different releases [7, 11]. These databases contain large-size non-homologous sequences, i.e., NRBD30 contains more than 125,000 sequences, NRBD60 contains 250,000 sequences, NRBD90 contains more than 500,000 ones. In addition to the size, these databases are essentially distinct in the composition of sequences. Analyzing amino acid sequences considering these bases, we calculated occurrence probabilities for different pairs of amino acid residues (P^k) at fixed distances k (the number of amino acid residues between them) (see Fig. 1).

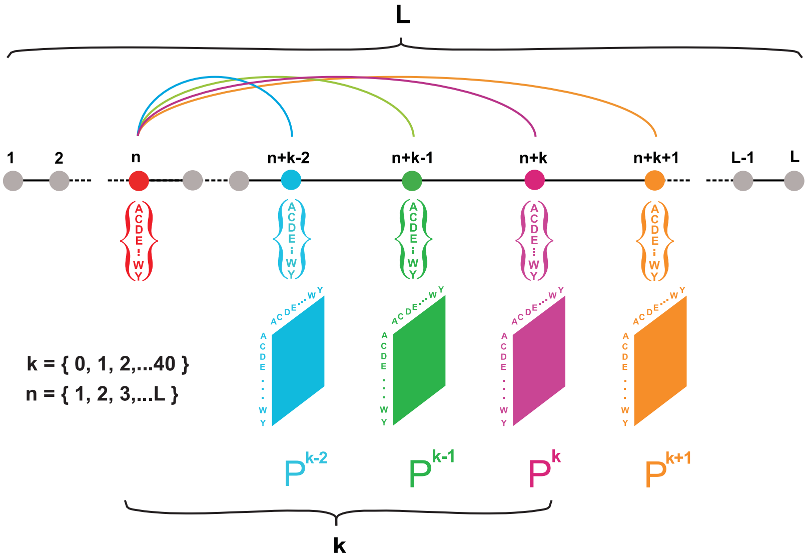

Figure 1. The scheme of analysis of sets of protein sequences in determination of frequency matrices for occurrence of amino acid pars of residues positioned at the fixed distance k in the sequences.

Forming the matrix P^k, we considered all sequences of a specific NRDB database release. Forty matrices of dimension 20×20 were formed. Shannon equation (1.1) [24] was used to calculate the amount of the information entropy for each matrix. The obtained values of the normalized information entropy S^k/S^0 as a function of the distance k are shown in Fig. 2 for the three used releases of NRDB database:

$$S^k = -\sum_{i=1}^{20}\sum_{j=1}^{20} P_{ij}^k \log_2 P_{ij}^k. \tag{1.1}$$

The value of the Shannon entropy depends on the size and composition of the studied database, therefore, we used the value S^0 as a normalization factor in order to neutralize the influence of the size of considered databases on the obtained functions. Figure 2 shows that the dependences obtained for all the three databases used here have an identical (S-shaped) character, and the positions of all local maxima and minima are completely the same. This allows to say that all obtained data characterize natural polypeptide chains in the whole and are not specific characteristics of a particular database. For $k > 30$ the dependence of $S^k/S^0(k)$ almost goes on a plateau, i.e., the sensitivity of the amino acid residues to their neighbours in the amino acid sequence is lost. This indicates that structural elements of natural polypeptide chains whose stability if determined by primary structure [1] should have the characteristic size of $\backsim$ 60 amino acid residues, which is in a good correspondence with the lower bound for the size of structural domains in known spatial structures of proteins. In addition, it is necessary to note that with the growth of k we can observe a decline in the amplitude of oscillations of the normalized information entropy S^k/S^0.

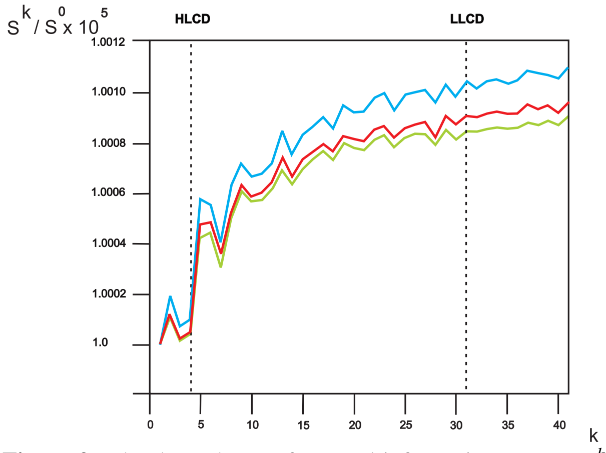

Figure 2. The dependence of normed information entropy S^k/S^0 on the distance k between amino acid sequences.

The choice of pentapeptide as a basic structural unit in the study of the amino acid sequences of proteins is caused by two reasons. The first is the observed sharp growth of the normalized Shannon entropy (see Fig. 2) for distances between the amino acid residues exceeding 5, i.e., $k > 5$. Another limiting factor is that the growth of the size of the considered fragment essentially decreases the frequency of its occurrence in the database of sequences, even in such a large one as NRDB. In addition, the testing of the algorithm we had performed showed that $k = 5$ gives the most qualitative detection of the hierarchical organization of the structural information in primary protein structures, although the hierarchical organization is also detected for other values of k. Below we use the term ‘information unit’ to indicate five-term fragments in primary structures of proteins.

Figure 2 clearly shows oscillations in the obtained dependence. Fourier analysis revealed two oscillations with the correlation periods of 3.6 and 2.9 amino acid residues in the information entropy. These periods correspond to the periods in stable α-spirals and spirals 3_{10} stabilized by hydrogen bonds.

1.2. Technique for analysis of protein information structure

The main idea allowing us to develop a new method of analysis of primary protein structure was proposed in [14, 15] and is that short (five-term) fragments of a polypeptide chain, i.e., information units (IU), are used as a unit of the protein sequence. The primary protein structure is considered in this case as a system of overlapping IU.

Let the primary structure of the protein $L = \{l_i\}$ of length N be a sequence of amino acids $l_i, i = 1, 2, \ldots, N$, where amino acids can be of 20 different types. Take a certain database consisting of non-homologous proteins (NRDB). Associate any subsequence of amino acids $K = k_1, \ldots, k_M$ of length $M = 5$ from sequence L with the frequency $f(K)$ of its occurrence in all sequences in the NRDB database.

Now compose the set of subsequences K' differing from K by replacement of one amino acid. Associate the sequences K' with the corresponding occurring frequencies $f(K')$ in the NRDB database. Sum the corresponding frequencies of occurrence $f(K')$ and obtain the function of occurrence of equivalent information units:

$$F(K) = f(K) + \sum_{K'} f(K'). \tag{1.2}$$

Now, given the considered protein with the sequence $L = \{l_i\}$ of length N, we consider all possible subsequences of length $M = 5$ of overlapping information units. A protein with the length of N amino acid residues contains $N - M + 1 = N - 4$ such subsequences for $M = 5$. Introduce the enumeration of sequences $K = K_i$ of length 5 according to their centers i, therefore, $i = 3, \ldots, N-2$, and consider the function of 'population' of a given position in the protein by equivalent information units

$$F(i) = \sum_{i-2}^{i+2} F(K_i) \tag{1.3}$$

which is the sum of frequencies of occurrences for five information units including this amino acid residue with the number i in this protein. The function $F(i)$ of 'population' of a given position in the protein by equivalent information units was constructed. Now proceed from the discrete function $F(i)$ to the continuous function $F(x)$ representing the histogram of $F(i)$. Further we applied a nonlinear smoothing to $F(x)$ (see Fig. 3). Now we construct the function $G(a,x)$ of nonlinear smoothing according to the following rule. Consider the smoothing function $\varphi(x)$ which is a continuous function with the support on the segment $[-1/2, 1/2]$, $\varphi(-1/2) = \varphi(1/2) = 0$, $\varphi(0) = 1$, $\varphi(x)$ is positive on the interval $(-1/2, 1/2)$, monotonically increasing on $[-1/2, 0]$, monotonically decreasing on $[0, 1/2]$, the graph of the function is symmetric relative to reflection with respect to the straight line $x = 0$. We assume that the function $\varphi(x)$ is smooth and its derivative does not turn to zero on the intervals $(-1/2, 0)$ and (0,1/2). Thus, the smoothing function can be taken as a certain central-symmetric function determined on some segment. In this paper we use a shifted and stretched Gaussian function given of a bounded interval.

We also consider shifts and stretches of the smoothing function

$$\varphi^{(a,b)}(x) = \varphi\Big(\frac{x-b}{a}\Big) \tag{1.4}$$

where $a \geqslant 1$. The function $\varphi^{(a,b)}$ has the support in the segment $[-a/2 + ba, a/2 + ba]$. Now define the function of nonlinear smoothing $G(x,a)$ for the function $F(i)$ determined on the discrete set of i by the following formula:

$$G(b,a) = \sup c, \quad c : \; c\varphi^{(a,b)}(x) \leqslant F(x) \;\; \forall x. \tag{1.5}$$

Thus, $G(b,a)$ is the maximal height *sup c* of the smoothing function of width a with the center of support at the point b such that it can be inscribed into the function

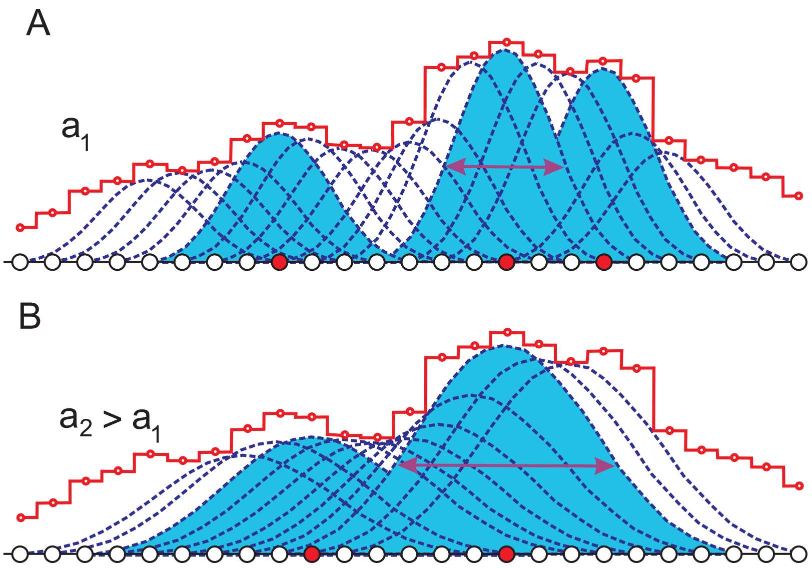

Figure 3. Nonlinear smoothing of the function of 'population' by equivalent information units for a given position in a protein.

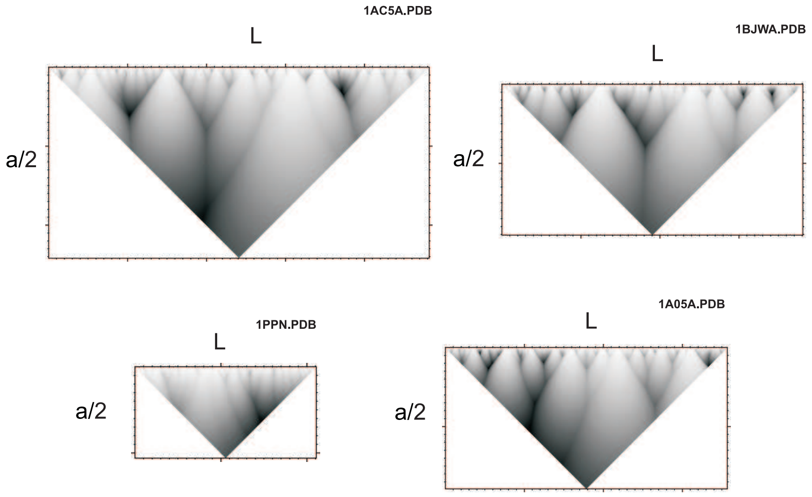

Figure 4. Examples of information structure calculations of various proteins. The axes correspond to the numbers of residues of the sequence *L* and the scale of smoothing *a*. Darker domains correspond to the centers of domains of size *a* with the maximal level of coordination between information units in a protein sequence. Hierarchically organized structures are seen in the figures.

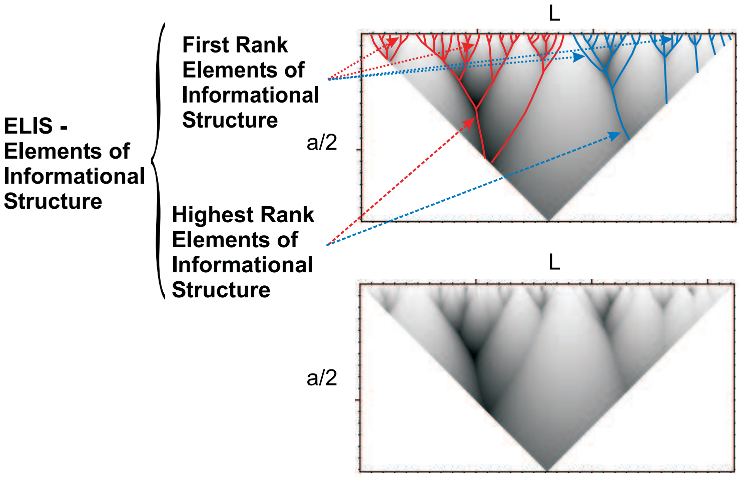

Figure 5. Construction of the graph of information structure indicating Information Structure ELements (ISEL) of different levels of hierarchy.

$F(x)$. The parameter a is called the scale of smoothing. The support of the function $G(x,a)$ has the following form. The function $G(x,a)$ can be distinct from zero for $a \in [1, N-4], x \in [2+a/2, N-2-a/2]$. Thus, the function of nonlinear smoothing has the support being a subset of the triangle with the vertices $(N/2, N-4), (2+1/2, 1), (N-2-1/2, 1)$ on the plane with the coordinates (x,a) (see Fig. 4). The set of values of the smoothing function $G(x,a)$ is called the information structure of the studied protein. Let $G(x,a)$ be the smoothing function described above. Construct all local maxima of the function $G(x,a)$ on the plane (x,a) with respect to x for all possible fixed a. If we replace the point of confluence of local maxima lines by the vertices and the lines joining them by edges, then we obtain the graph of information structure (see Fig. 5). The hierarchical structure obtained by the method presented here in the amino acid sequence of the protein reflects the occurrence of IU in databases of the protein sequences and interconnections between them.

2. Results

2.1. Analysis of information structure as a method of protein engineering

The above method of Information Structure Analysis of sequences (ISAN) has allowed to represent the primary protein structure as a system of hierarchically organized elements for the first time. We compared the detected hierarchically organized elements with the structural elements of proteins used in classic views [15]. Figure 6 shows that ISEL of the highest rank correspond to structural domains in proteins with clearly expressed domain structure. It should be noted that such correspondence does not depend on elements of the secondary structure of a protein forming structural domains.

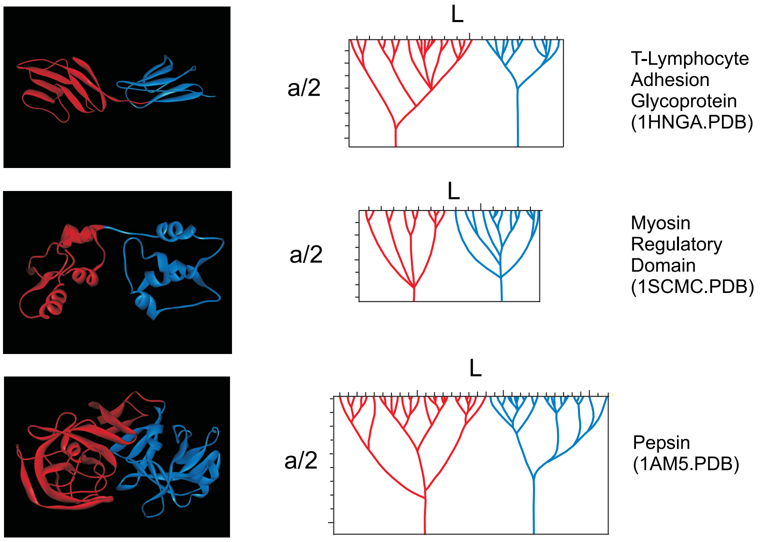

Figure 6. Correspondence of structural domains and ISEL of the highest rank in proteins. The right side presents information structures of proteins where different colours highlight ISEL of the highest rank. L is the ordinal number of the amino acid residue in the protein sequence, a is the smoothing scale. In the information structure, the green dotted line bounds the domain of definition of the function $G(x,a)$. The left side shows spatial structure of proteins where the local colour of the polypeptide chain corresponds to the colour of ISEL of the highest rank.

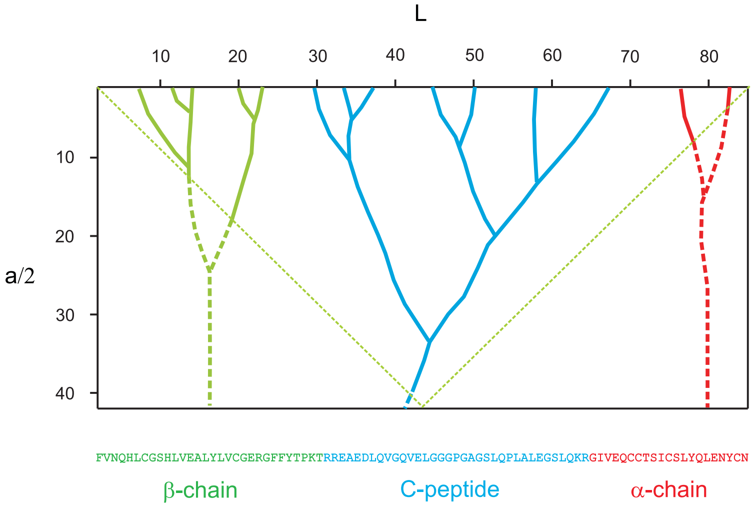

Figure 7. The information structure of proinsulin where different colours mark ISEL of the highest rank. L is the ordinal number of an amino acid residue in the protein sequence, a is the scale of smoothing. The green dotted line bounds the domain of definition of the function $G(x,a)$ in the information structure. In the protein sequence, different colours indicate the fragments of sequence corresponding to particular ISEL of the highest rank. The β-chain is painted green and the α-chain of insulin is painted red. The blue colour indicates the C-peptide chipped off the native structure in the hormone maturation process.

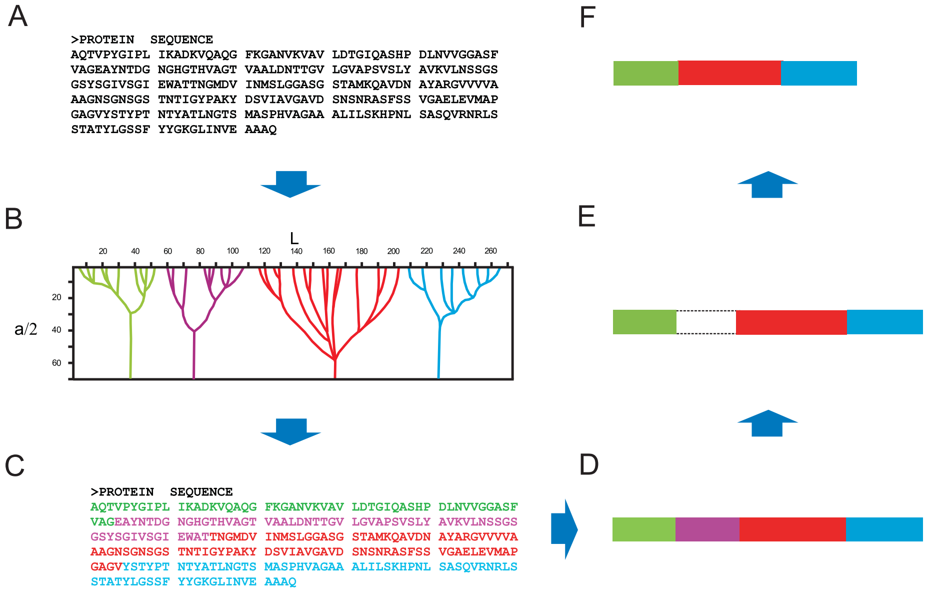
A
>PROTEIN SEQUENCE
AQTVPYGIPL IKADKVQAQG FKGANVKVAV LDTGIQASHP DLNVVGGASF
VAGEAYNTDG NGHGTHVAGT VAALDNTTGV LGVAPSVSLY AVKVLNSSGS
GSYSGIVSGI EWATTNGMDV INMSLGGASG STAMKQAVDN AYARGVVVVA
AAGNSGNSGS TNTIGYPAKY DSVIAVGAVD SNSNRASFSS VGAELEVMAP
GAGVYSTYPT NTYATLNGTS MASPHVAGAA ALILSKHPNL SASQVRNRLS
STATYLGSSF YYGKGLINVE AAAQ
B
L
20 40 60 80 100 120 140 160 180 200 220 240 260
a/2
20 40 60
C
>PROTEIN SEQUENCE
AQTVPYGIPL IKADKVQAQG FKGANVKVAV LDTGIQASHP DLNVVGGASF
VAGEAYNTDG NGHGTHVAGT VAALDNTTGV LGVAPSVSLY AVKVLNSSGS
GSYSGIVSGI EWATTNGMDV INMSLGGASG STAMKQAVDN AYARGVVVVA
AAGNSGNSGS TNTIGYPAKY DSVIAVGAVD SNSNRASFSS VGAELEVMAP
GAGVYSTYPT NTYATLNGTS MASPHVAGAA ALILSKHPNL SASQVRNRLS
STATYLGSSF YYGKGLINVE AAAQ
D
E
F

Figure 8. The scheme of the method for obtaining new recombinant proteins on the base of the natural protein sequences. A) The primary structure of the original natural protein. B) The information structure of the original protein where different colours indicate ISEL of the highest rank. *L* is the ordinal number of an amino acid residue in the protein sequence, *a* is the scale of smoothing. C) The sequence of the original natural protein where different colours indicate fragments corresponding to ISEL of the highest rank of information structure. D) The linear model of the original protein sequence where different colours indicate fragments corresponding to ISEL of the highest rank. E) The linear model of the original protein sequence where different colours indicate fragments corresponding to ISEL of the highest rank and white colour indicates the fragment marked for deletion. F) Linear model of a new protein sequence where highlighted fragments correspond to ISEL of the highest rank.

It is known that the structural domains are stable structural elements and can form spatial structure by themselves. The correspondence of the highest rank ISEL to structural domains can indicate the important role that those ISEL play in the formation of spatial structure of proteins. In addition, one can assume that the elements of 3D structure corresponding to the elements of information structure of all other ranks are also stable elements of the 3D protein structure. An additional reason confirming this point of view is the information structure of proinsulin. Figure 7 presents the sequence and information structure of the predecessor of insulin. This protein has no visible domain structure. A fragment of a sequence called C-peptide is chipped off from proinsulin in the process of maturing. A specific ISEL of the highest rank corresponds to C-peptide in the information structure.

All these data allow us to suggest that one can remove the fragments corresponding to particular ISEL of the highest rank from a natural polypeptide sequence not violating the folding mechanism and structural stability of other elements of the 3D protein structure. Figure 8 presents the scheme of the method proposed here and allowing one to obtain amino acid sequences of new recombinant proteins based on the natural protein sequences.

The ISAN method described above was used in experimental studies of proteins in [10, 16, 17, 22, 25].

3. Discussion

In this paper we study the regularities of the structural organization of information stored in the primary structure of proteins. It was shown that the form of dependence of the Shannon entropy as a function of the distance between amino acid residues is invariant for different sets of protein sequences. The existence of long-range sensitivity between a.a.r. was revealed in the primary protein structure and a constant and low value of positional information entropy was detected for distances between a.a.r. less than six positions. The data obtained here allowed us to propose a new method of description of protein sequences where an elementary element of a sequence is the 'information unit' being a group of several adjacent a.a.r. The choice of the size of 'information unit' is determined by two experimental facts. First, the graph of the Shannon entropy calculated from matrices of occurrence of amino acid residue pairs as a function of the distance between residues k for $k > 5$ indicates a sharp growth of its value. Second, with the increase of the size of 'information unit' the occurence probability decreases. A limiting factor is that the increase of its occurrence in the base of homologous protein sequences, even in such a large one as NRDB.

The new method of primary protein structure coding allowed us to develop the ISAN method thus presenting an ability to detect a hierarchical organization of structure information stored in the amino acid sequences of proteins. The information structure of protein is a set of hierarchically organized ISEL.

The ISAN method developed here is a variant of wavelet analysis. Technically, the method is applicable to the amino acid sequences with the minimum length of more than 5 a.a.r., however, the ISAN method gives significant scientific results when the length of the studied amino acid sequence is at least several dozens of a.a.r. It was shown that ISEL of the highest rank corresponds to structural domains of proteins in proteins with clearly expressed domain structure. A new approach to design of sequences of new recombinant proteins is proposed.

References

1. C. B. Anfinsen, E. Haber, M. Sela, and F. H. White Jr., The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. USA* **47** (1961), 1309–1314.

2. I. N. Berezovsky and E. N. Trifonov, Loop fold structure of proteins: resolution of Levinthal's paradox. *J. Biomol. Struct. Dynam.* **20** (2002), No. 1, 5–6.

3. A. C. Camproux, R. Gautier, and P. Tuffery, A hidden Markov model derived structural alphabet for proteins. *J. Mol. Biol.* **339** (2004), 591–605.

4. K. A. Dill, From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4** (1997), No. 1, 10–19.

5. A. V. Efimov, Structural trees for proteins containing phi-motifs. *Biochemistry (Moscow)* **73** (2008), No. 1, 23–28.

6. S. Henikoff and J. G. Henikoff, Automated assembly of protein blocks for database searching. *Nucleic Acids Research* **19** (1991), No. 23, 6565–6572.

7. L. Holm and C. Sander, Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* **14** (1998), No. 5, 423–429.

8. B. Honig, Protein folding: from the Levinthal paradox to structure prediction. *J. Mol. Biol.* **293** (1999), No. 2, 283–293.

9. M. Karplus, The Levinthal paradox: yesterday and today. *Fold. Des.* **2** (1997), No. 4, 569–575.

10. E. I. Kovalenko, L. M. Kanevskii, A. M. Sapozhnikov, and A. N. Nekrasov, Application of protein information structure analysis for detection of 70 kDa heat shock protein sections activating NK-cells. In: *Stochastic and Computer Simulation of Systems and Processes*. Ya. Kupala Grodno State. Univ., Grodno, 2011, ISBN–978-985-515-495-3, pp. 379–384.

11. W. Li, L. Jaroszewski, and A. Godzik, Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* **18** (2002), No. 1, 77–82.

12. C. Micheletti, F. Seno, and A. Maritan, Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *PROTEINS: Structure, Function, and Genetics* **40** (2000), 662–674.

13. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247** (1995), No. 4, 536–540.

14. A. N. Nekrasov, Entropy of protein sequences: an integral approach. *J. Biomol. Struct. Dyn.* **20** (2002), No. 1, 87–92.

15. A. N. Nekrasov, Analysis of the information structure of protein sequences: a new method for analyzing the domain organization of proteins. *J. Biomol. Struct. Dyn.* **21** (2004), No. 5, 615–624.

16. A. N. Nekrasov, L. E. Petrovskaya, V. A. Toporova, E. A. Kryukova, A. V. Rodina, E. Y. Moskaleva, and M. P. Kirpichnikov, Design of a novel interleukin-13 antagonist from analysis of informational structure. *Biochemistry (Moscow)* **74** (2009), No. 4, 399–405.

17. A. N. Nekrasov, V. V. Radchenko, T. M. Shuvaeva, V. I. Novoselov, E. E. Fesenko, et al., The novel approach to the protein design: active truncated forms of human 1-Cys peroxiredoxin. *J. Biomol. Struct. Dyn.* **24** (2007), 455–462.

18. C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, CATH — a hierarchic classification of protein domain structures. *Structure* **5** (1997), No. 8, 1093–1108.

19. L. Pauling, R. B. Corey, and H. R. Branson, The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA* **37** (1951), 205–211.

20. V. V. Poroikov, N. G. Esipova, and V. G. Tumanyan, Distribution of amino acid residues in primary protein structure. *Mol. Biol.* **18** (1984), No. 2, 541–547.

21. O. B. Ptitsyn and M. V. Volkenstein, Protein structure and neutral theory of evolution. *J. Biomol. Struct. Dyn.* **4** (1986), 137–156.

22. V. V. Radchenko, A. N. Nekrasov, T. M. Shuvaeva, M. I. Merkulova, and V. M. Lipkin, Analysis of informational structure of protein amino acid sequences and obtaining large biologically active peptides of human 1-Cys peroxiredoxin. *J. Peptide Sci.* **10** (2004), No. 8, 258.

23. J. S. Richardson, The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34** (1981), 167–339.

24. C. E. Shannon, A mathematical theory of communication. *Bell System Tech. J.* **27** (1948), 379–423.

25. L. N. Shingarova, L. E. Petrovskaia, A. N. Nekrasov, E. A. Kriukova, E. F. Boldyreva, S. A. Iakimov, S. V. Gur'ianova, D. A. Dolgikh, and M. P. Kirpichnikov, Production and properties of human tumour necrosis factor peptide fragments. *Bioorg Khim.* **36** (2010), No. 3, 327–336.

26. G. Szoniec and M. J. Ogorzalek, Entropy of never born protein sequences. *SpringerPlus* **2** (2013), No. 1, 200.

27. O. Weiss, M. A. Jimenez-Montano, and H. Herzel, Information content of protein sequences. *J. Theor. Biol.* **206** (2000), No. 3, 379–386.