Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2021 doi:10.20944/preprints202106.0661.v1

Article

Hierarchical structure of protein sequence

Alexei N. Nekrasov [1], Yuri P. Kozmin [1], Sergey V. Kozyrev [2], Rustam H. Ziganshin [1], Alexandre G. de Brevern [3], Anastasia A. Anashkina [4,*]

[1] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, The Russian Academy of Sciences, Miklukho-Maklaya St. 16/10, 117997, Moscow, Russia; alexei_nekrasov@mail.ru, zibotic@mail.ru, ziganshin@mail.ru
[2] Steklov Mathematical Institute of Russian Academy of Sciences, 8 Gubkina St. Moscow, 119991, Russia; kozyrev@mi.ras.ru
[3] INSERM UMR S-1134, DSIMB, Univ. Paris, Univ de la Reunion, Univ des Antilles, INTS, lab. of excellence GR-Ex 6, rue Alexandre Cabanel 75739 Paris Cedex 15, France2; alexandre.debrevern@gmail.com
[4] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilov St. 32, 119991 Moscow, Russia; nastya@eimb.ru

* Correspondence: nastya@eimb.ru

Abstract: Most non-infectious diseases are associated with dysfunction of proteins or protein complexes. Association between sequence and structure is analyzed since a long time, and analysis of sequence organization in domains and motifs is actual research area. A mathematical method is proposed here to identify the hierarchical organization of protein sequences. The method is based on pentapeptide as a unit of protein sequences. This method was applied on a non-homologous dataset of protein sequences. The analysis revealed 11 hierarchical levels of protein sequence organization, showing the relationship of these multiple fragments of sequences. Using different examples, we illustrated how the fragments of the spatial structure of protein correspond to the elements of the hierarchical structure of the protein sequence. A hierarchical structure is observed in the protein sequence. This methodology is an interesting basis for mathematically based classification of elements of spatial organization of proteins. Elements of the hierarchical structure of different levels of the hierarchy can be used for biotechnological and medical problems.

Keywords: protein structure; hierarchy; protein sequence; ANIS method; supersecondary structure.

1. Introduction

The multiplicity of functional characteristics of proteins is associated with a wide variety of their spatial structures. This diversity is ensured by the different arrangement of amino acid residues in their sequences. Hierarchical organization in the spatial structure of proteins was revealed by using various approaches, for example, such as calculating the local packing density of atoms in the structure [1–4], interaction energy inside a protein globule [5], Fuzzy Oil Drop sites [6] or hydrophobic folding nuclei [7]. Hierarchical elements of proteins were considered as elements of protein folding at different stages [8–11]. The analysis of protein sequences had been made at the level of protein sequence domains with the Pfam database [12]. Translation of protein functional annotation is then done by simple similarity measure. At a lower level, some repetitive elements (patterns) were often observed [13–15]. The patterns were used to predict structure and function of proteins [16–18]. However, in contrast to the structural elements obtained in the analysis of spatial structures, protein sequences have not previously been found to have a hierarchical structure.

In this paper, we consider proteins as a system of hierarchically organized elements. In the spatial structure, elements at the top level of the hierarchy are structural domains[1,9]. Structural domains have almost all the features of natural polypeptide chains and, first of all, the ability to self-assemble. Obviously, these features should be

© 2021 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2021 doi:10.20944/preprints202106.0661.v1

visible in the protein sequences. Early studies of protein sequences showed that the complexity of the primary structures of natural proteins differs from random polypeptides of the same amino acid composition by about 1% [19]. This lead to the representation that the protein sequences are "slightly edited random sequences" [20,21]. Obviously, such concepts do not correspond to the observed structural and functional properties of proteins as seen with the large number of approaches to predicting coding zones in genomes by Hidden Markov Model approaches[22]. To resolve this contradiction, a new paradigm was proposed by us in [23–25], which was confirmed in a number of experimental researches [26–28]. We showed [23] that the lowest level of Shannon entropy [29] is observed inside blocks of five amino acid residues in protein sequences. Taking a fragment of five amino acid residues as a unit of protein sequence, we proposed the ANIS (analysis of Informational structure) method for identifying hierarchically organized structures in protein sequences [25,30]. This method identifies tree-like hierarchical structures (graphs) in a protein sequence. Several examples shown that free-standing graphs correspond to structural domains [26]. Subsequently, the revealed hierarchical elements were used by us for protein design [31–34] and for study of the mechanisms of protein function [26–28]. It was experimentally shown in [28,31,33,34] that the removal of sequence fragments corresponding to free-standing graphs from the native protein sequence leads to minimal folding distortion of the recombinant protein. In [26,28], using the analysis of hierarchical elements in the protein structure, mechanical models of functioning of protein molecular machines were proposed.

An approach in which a protein sequence is considered as a system of blocks of five amino acid residues was used in the literature to create a protein structural alphabet [35], to study topologically stable elements of protein spatial structure of the lowest level [36], to describe the folding of protein molecules [37,38]. In 2020, Kaushik and Zhang showed that the use of tripeptides makes it possible to differentiate natural protein sequences and artificial polypeptide sequences that do not exist in nature [39]. We have shown earlier [23], that low level of informational entropy is observed in a range from 2 to 8 amino acids fragment; the lowest level is observed for pentapeptides. We assume that if a similar work uses pentapeptides instead of tripeptides, the results will be even more revealing.

Application of the ANIS method to a large number of natural protein sequences has led us to observe that there are characteristic sizes of sequence fragments in which tree-like graphs are divided into smaller hierarchical elements. Size of the fragments are smaller than structural domains, but larger than the structural alphabet proposed by de Brevern [35]. The present work is devoted to the analysis of the sizes of the revealed hierarchically organized elements of the protein sequence.

2. Results

2.1. Correllation between ELIS of different hierarchical levels

The correlation matrix $(r_{yy'})$ contains the data on correlations between ELIS of different values of y and y' in protein hierarchical structure (eq. (9)). Correlation matrix $(r_{yy'})$ was built for 24 647 protein sequences with sizes in between 50 to 400 of amino acid residues from the NRDB90 data base [49].

Figure 1 (A-I) shows a graphical representation of the matrix elements $r_{yy'}$ calculated by eq. (9). This matrix reflects agreements of branching points for different y values. Values of matrix elements of $(r_{yy'})$ lie in the interval from 0 to 1. Matrix elements equal to 1 are shown in black and located only on the diagonal (y = y′). Other matrix elements are displayed in the figure in gray if they exceed the specified threshold value. The threshold value is shown above the corresponding figure.

If branching points in hierarchical structures are situated at close y values then matrix $(r_{yy'})$ will contain nonzero elements around the diagonal (see Figure 4). These non-zero elements (the red squares at the figure) form continuous square areas (CSA).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2021 doi:10.20944/preprints202106.0661.v1

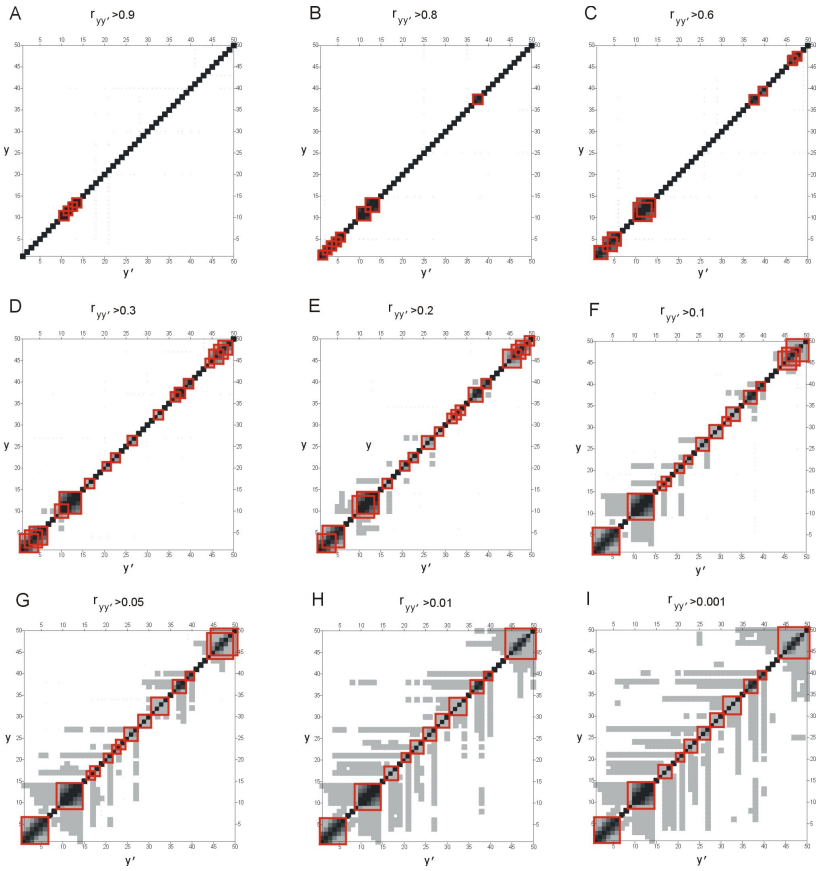

Figure 1. Graphical representation of the matrix $r_{yy'}$. Matrix elements are displayed in the figure if they exceed the specified threshold value. Nine threshold values were used from 0.9 to 0.001 (A-I). All matrix elements which values $r_{yy'}$ exceed threshold were displayed in gray. Matrix elements with values $r_{yy'}$ equal to 1 are displayed in black. Matrix elements $r_{yy'}$ (given by eq. (9)) reflect correlation of vectors V_y of first derivatives $\tilde{S'}_I(y)$ for different thresholds.

Let's look at how the gray elements change with decreasing of threshold value. If there is a region of neighbor gray cells, then a CSA is formed in the diagonal region. It is highlighted in red. Let's reduce the threshold value. The number of gray cells will increase and they will form CSA in different places and of a greater size. At 0.01 of threshold value (Fig. 1H) comes a moment that new gray cells do not appear, new CSA in the diagonal area do not formed. Thus, the structure of CSAs in the diagonal region became stable.

Table 1 shows the sizes and positions of the stable CSA. For the set of samples of non-homologous protein sequences analyzed in this work, these near-diagonal CSA correspond to the levels of hierarchical organization in the protein sequences, and, therefore, in the spatial structure of proteins. We assume that there is a relationship between the levels of organization in the protein sequence and its spatial structure.

Table 1. Levels of hierarchical organization of protein sequence elements.

Level of structural organization	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
Range of values of semi width of smoothing function	1-6	9-14	16-18	20-21	22-24	25-27	28-30	31-34	36-38	39-40	44-50

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2021 doi:10.20944/preprints202106.0661.v1

2.2.Hierarchy in the spatial structure of proteins

The example below illustrates the relationship between hierarchical elements of the sequence and the corresponding fragments of the spatial structure. Figure 2A shows the hierarchical structure of the protein sequence of the photosynthetic reaction center from *Rhodopseudomonas viridis* (PDB id 1PRC, chain C) obtained using the ANIS method.

The architecture of N- and C-terminal ELISes can be considered qualitatively if they are approximated beyond the boundaries of the triangular domain of the function $H_1(x, y)$ definition. Continue the branches in the same direction when they reach the border of the triangular area. It can be seen that the ELIS located at the N- and C-ends of the sequence do not tend to merge with the ELIS located in the center of the sequence. Therefore, we can assume that the elements of the spatial structure at the N- and C- ends of the sequence are weakly correlated with the ELIS in the central part of the protein.

In the central region of the sequence, between positions 90 and 248, two independent ELIS are defined (green and red) (Figure 2A). The boundary between them is the residue 121. It should be noted that ELIS, formed between 121 and 248 residues (red color), has a very complex structure; the branches forming it merge at **y** values of more than 100. Figure 2A shows the spatial structural elements of the protein corresponding to ELIS also. First ELIS (green color) corresponds to α-helix with loop-shaped structures at its ends. Second ELIS, highlighted in red, has complex spatial structure.

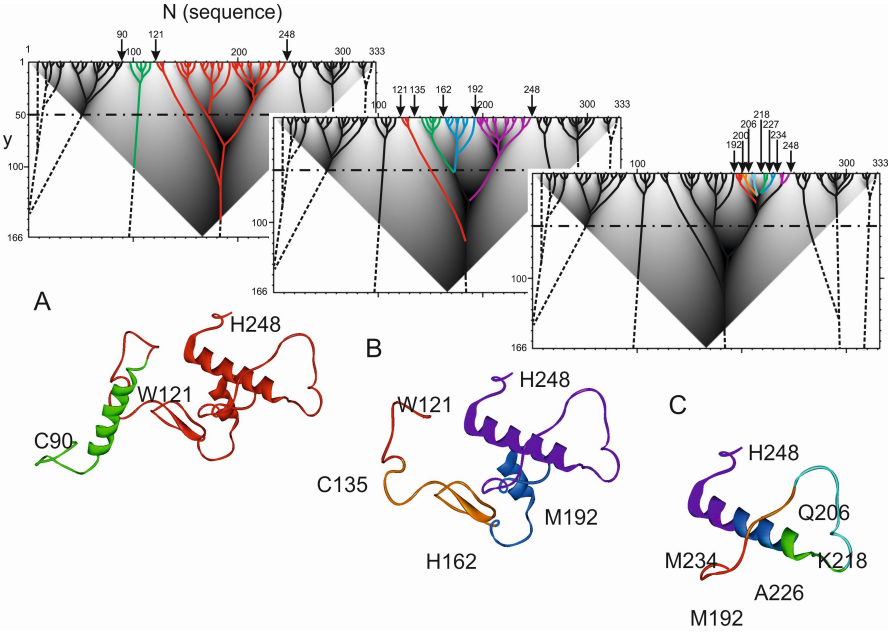

Figure 2. The correspondence between the hierarchical structure of the sequence and the spatial structure of photosynthetic reaction center protein from Rhodopseudomonas viridis fragments (PDB id 1PRC chain C). Dot-dash line marks y=50. Dashed lines show the ELIS approximation beyond the boundary of the triangular definition area. (A) Hierarchical structure of the protein sequence with two colored high-rank ELISes: 90-121 (green) and 121-248 (red) with corresponding spatial structure. ELISes located at the N- and C-ends of the protein sequence are shown in black on the hierarchical structure. (B) Elements of middle-level hierarchical structure are shown in red (121-135), in green (135-162), in blue (162-192) and in violet (192-248) with corresponding spatial structure. (C) Elements of low-level hierarchical structure are shown in red (192-200), in orange (200-206), in cyan (206-218), in green (218-227), in blue (227-234) and in violet (234-248) with corresponding spatial structure.

Next, consider spatial structure that corresponds to ELIS 121-248 (Fig. 2B). The first element 121-135 is a small loop-like structure that terminates the α-helix and provides turn of the polypeptide chain (marked in red in Figure 2B). The second element 135-162 forms the β-hairpin and N- and C-turns of adjacent polypeptide chain (marked green in the Figure 2B). The third element 162-192, marked blue, is a short α-helix with N- and C- turns of adjacent polypeptide chain. The last ELIS 192-248 is highlighted violet (Fig.2B) and is the longest at this level of the hierarchy. Its spatial structure is quite complex and includes an α-helix and an irregular fragment of the polypeptide chain.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2021 doi:10.20944/preprints202106.0661.v1

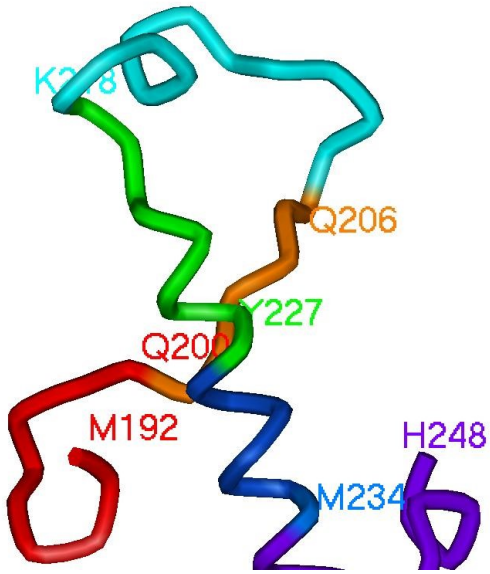

Figure 3. Division into elements at a lower level of the hierarchy the ELIS 192-248. Red (192-200), orange (200-206), blue(206-218), green (218-227), blue (227-234) and violet (234-248) elements of the hierarchical information structure are marked.

Spatial structure of 192-248 ELIS is quite complex (Figure 2C, Figure 3). It consists of 6 lower-level ELIS, which are: red ELIS (192-200) – a fragment of the polypeptide chain terminates the α-helix and implements the turn of the polypeptide chain; orange ELIS (200-206) – a fragment of the polypeptide chain located in an extended conformation; cyan ELIS (206-218) is a fragment of a polypeptide chain that forms a Π-shaped loop with a helix turn located in the C-terminal part of the fragment; green ELIS (218-227) – a fragment of the polypeptide chain that forms the initiator of the α-helix; blue ELIS (227-234) is a fragment of the polypeptide chain that supports the α-helix conformation; violet ELIS (234-248) is a fragment of a polypeptide chain that sequentially forms the termination of a α-helix, reverse rotation, and initiation of the next α-helix.

3. Discussion

Analyzing hierarchical structures of proteins, we can summarize, that ELIS of the upper levels of the hierarchy usually correspond to structural domains [35]. At the lower levels of the ELIS hierarchy, elements of the super-factor structure (β- hairpins, α- hairpins), elements of the secondary structure (α-helixes, β - strands, or elongated structures), as well as spatial elements that provide a transition from one known element of the spatial organization to another are found. A distinctive feature of the ANIS method is the recognition of irregular, "non-classic" loop-liked fragments of the polypeptide chain in the identified structural elements.

Structures revealed using the ANIS method are the result of the molecular evolution of polypeptide chains, since the elements of the lower level of the hierarchy are formed by more common pentapeptides[50,51] (see Fragments and hierarchical structure of a protein, eq. 1). That is, not only the regular structural elements themselves in the spatial structures of proteins are important, but also the direction of the further course of the polypeptide chain. Undoubtedly, this plays a very important role in the folding of the polypeptide chain. Note that the proposed method for identifying structural elements allows us to reveal new classes of structural elements. Note, some of the elements detected by ANIS method do not correspond to usual structure classification elements like α-helices and β-sheets (see description for Figure 5C, 6 in the text).

4. Materials and Methods

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2021 doi:10.20944/preprints202106.0661.v1

4.1. Fragments and hierarchical structure of a protein

As mentioned earlier, the greatest self-consistency is observed within blocks of five amino acid residues [32]. Let us consider fragments of length five in proteins. We consider all possible overlapping fragments, i.e. neighboring fragments overlap with four residues. To any fragment A of length five from the protein sequence we put in correspondence the frequency – the number $\varphi(A)$ of occurrence of the fragment in the database of non-homological protein sequences. We will consider the total frequency:

$$\Phi(A) = \sum_{J:d(A,J)\le\delta} \varphi(J) \qquad (1)$$

over sequences J of length five with Hamming distance from A not larger than $\delta = 1$. In our papers [34,39] Hamming distance δ equal to one was used also, i.e. we summarized over fragments which differ from the initial fragment at no more than one amino acid residue.

We will put in correspondence to a protein sequence (as a sequence $I = i_1 ... i_N$ of amino acid residues) a sequence of fragments I_j (pentapeptides) enumerated by central residues in fragments, i.e. $j = 3,...,N-2$. Let us put in correspondence to a j-th pentapeptide in the protein I a value given by (1). Here we sum over pentapeptides which differ from I_j in no more than one residue (Hamming distance δ not larger than one) and frequencies are taken from the database NRDB90. Let us introduce for a protein with sequence I the function $f_I(j) = \Phi(I_j)$ of total frequency of pentapeptides.

Let us consider the Gaussian function on real axis (as in [34,39]):

$$g(x,y) = \frac{1}{y\sqrt{2\pi}} e^{-\frac{x^2}{2y^2}} \qquad (2)$$

and let us introduce the smoothing distribution of total frequency of pentapeptides in a protein I as a convolution of the Gaussian function and the total frequency of pentapeptides in a protein:

$$F_I(x) = \sum_{j=3}^{N-2} f_I(j) g(x-j,y) \qquad (3)$$

Then, let us fit Gaussian functions in the graph of the function $F_I(x)$, i.e. we will get the function:

$$H_I(x,y) = \max h : \min z : [F_I(z) - he^{-\frac{(z-x)^2}{2y^2}}] \ge 0 \qquad (4)$$

which measures the height of Gaussian function with center in x and width $2y$ which can be fit in the graph of function $F_I(x)$.

Function $H_I(x,y)$ (4) defined in isosceles triangle in the coordinate plane (x,y) with the base $x \in [1,N]$ at the abscissa axis and height $y \in [1, N/2]$. An example of the function $H_I(x,y)$ is shown at the figure 4A. Local maxima of the function $H_I(x,y)$ (with respect to the abscissa when the ordinate is fixed) constitute a tree-like graph (as shown in Figure 4B). It is natural to put in correspondence branches of this graph to elements of hierarchical structure of protein sequence (namely ELIS, hierarchical ELements of Informational Structure [39], see Figure 4).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2021 doi:10.20944/preprints202106.0661.v1

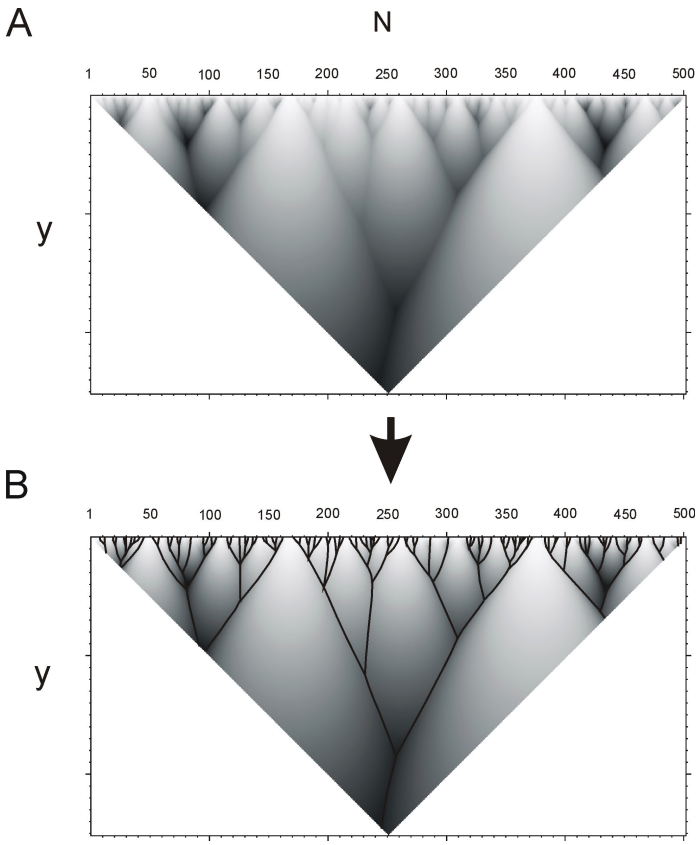

Figure 4. Example of hierarchical structure of protein sequence determined by function (4). (A). Hierarchical structure of the sequence of catalase protein (UniProt id P29422) obtained by the ANIS (analysis of hierarchical structure) method. (B). Tree-like graph is constructed using local maxima of the function $H_I(x,y)$ (4). ELIS are given by branches of the graph. Notations of axes: N - number of amino acid residue in the protein sequence, y – semi-width of the Gaussian function (2).

Each hierarchical element is characterized by position in the protein sequence, number of branches contained in the element and the rank in the hierarchy. Let us consider one of hierarchical elements of the graph of $H_i(x,y)$ with branching at the point $(x_0.y_0)$. At this point, the graph branches into several elements with lower levels of the hierarchy (Figure 4B).

Let us note that hierarchical structure of protein sequences can be very diverse. Figure 2 shows hierarchical structures for several protein sequences from UniProt database.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2021 doi:10.20944/preprints202106.0661.v1

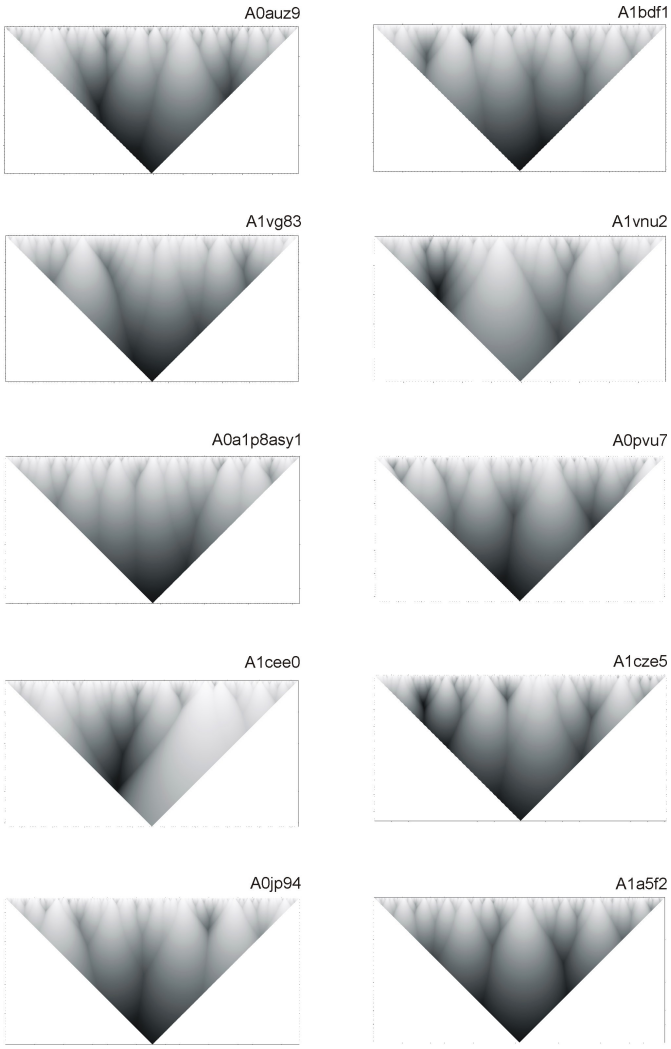

Figure 5. Hierarchical structures of some protein sequences. Codes of sequences from database UniProt are indicated above the pictures.

4.2. Hierarchy in structure of protein sequences

We selected non-redundant set of protein sequences from 50 to 400 amino acid residues from the NRDB90 database [49]. As a result of this selection, we obtained a set of 24,647 protein sequences. As it was shown in [32], a set of several thousand protein sequences is sufficient to reveal the regularities common to all proteins. The protein dataset must satisfy only two criteria: be large enough and not contain homologous sequences.

For each protein from this set let us compute function $H_I(x,y)$ given by eq. (4). Lengths **N** of protein sequences vary from 50 to 400.

For a protein I and fixed half-width y let us consider function $H_I^{norm}(x,y)=\dfrac{H_I(x,y)}{\int H_I(x,y)dx}$ as a probability distribution, i.e. let us normalize function $H_I(x,y)$ to satisfy

$$\int H_I^{norm}(x,y)dx=1 \qquad (5)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2021 doi:10.20944/preprints202106.0661.v1

Shannon entropy for this probability distribution has the form

$$S_I(y) = -\int H_I^{norm}(x,y)\log H_I^{norm}(x,y)dx \qquad (6)$$

For a partition of segment $[0,1]$ in n equal segments (with length $1/n$) entropy of partition will be equal

$$\mathbf{2} - \sum_{i=1}^{n} \frac{1}{n}\log\frac{1}{n} = -\log\frac{1}{n} = \log n \qquad (7)$$

In this case entropy depends only on number of fragments n.

Let us consider the following difference between real entropy of protein sequence (6) and model entropy of distribution (7)

$$\tilde{S}_I(y) = S_I(y) - \log n \qquad (8)$$

This function has leap in points of branching of the tree-like graph where ELIS of smaller rank emerges. Now let calculate the derivative $\tilde{S}'_I(y)$ of the function $\tilde{S}_I(y)$ with respect to y. Local maxima of this function will correspond to branching points of the graph. Figure 3B shows an example of hierarchical structure of a protein sequence and the derivative (Fig. 3A).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2021 doi:10.20944/preprints202106.0661.v1

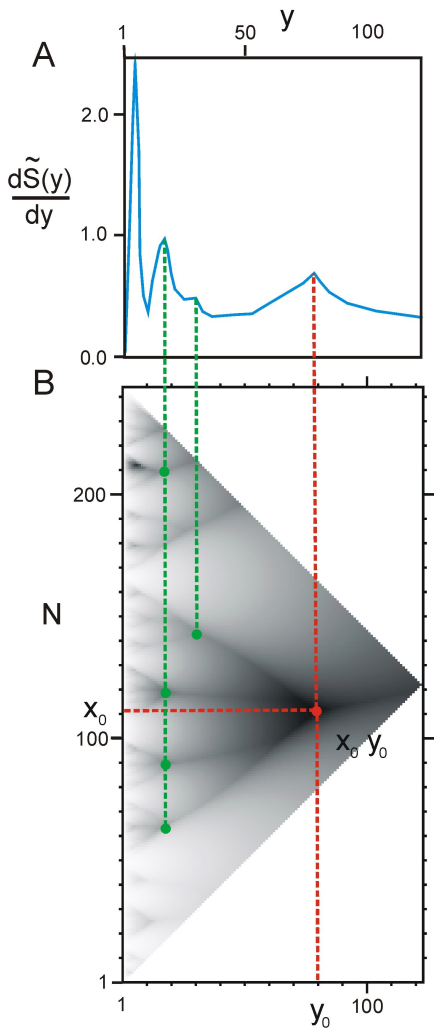

Figure 6. First derivative $\tilde{S}'_I(y)$ with respect to y. (A) Graph of the derivative of the difference between real and model entropy of partition (eq. (8)). Dashed line shows relation between points of branching in the hierarchical structure of protein sequence (Figure 6B) and maxima at the graph of the first derivative (Figure 6A). (B) Hierarchical structure of protein sequence of 26S proteasome regulatory subunit rpn10 (UniProt id O94444) where branching points are indicated. Branching point $(x_0.y_0)$ is mentioned above and indicated by red. Other branching points are indicated by green.

Next, let us consider for each fixed $y = 1,...,N$ a vector V_y which contains derivatives $\tilde{S}'_I(y)$. This vector consist of I elements $\tilde{S}'_I(y)$ from each protein sequence of 24 647 dataset. Let us consider the correlation of these vectors for different y values:

$$r_{yy'} = \frac{\langle V_y, V_{y'} \rangle}{\sqrt{\langle V_y, V_{y'} \rangle \langle V_{y'}, V_{y'} \rangle}}, \quad \langle A, B \rangle = \sum_I A_I B_I \tag{9}$$

5. Conclusions

In this paper, a strict description of the method for studying of information in protein sequences is given – the method for analyzing the hierarchical information structure of protein sequences (ANIS method) [20]. We applied

ANIS method to 24647 non-homologous protein sequences with sizes from 50 to 400 residues from the NRDB90 database [49] and identified elements that form a hierarchical structure in the protein sequence. We have shown that there is a correlation between the identified elements of different sizes. In this work, we revealed 11 levels of structural organization in proteins. In addition, we have shown the relationship between the identified elements of the sequence and the elements of the spatial structure of proteins.

The results obtained open up the possibility of creating a mathematically substantiated hierarchical classification of the structural elements of proteins. In the future, these results can also be used to study the molecular evolution of structural domains, the design of recombinant proteins, and the design of proteins with new types of spatial organization.

Author Contributions: conceptualization, A.N., A.A. and A.B.; methodology, Y.K., S.K. and A.N.; software, Y.K. and A.N.; validation, A.A., and Y.K.; formal analysis, Y.K., A.A. and. S.K.; investigation, A.N. and A.A.; resources, R.Z.; data curation, Y.K. and R.Z.; writing—original draft preparation, A.N. and A.A.; writing—review and editing, A.N., A.A. and A.B.; visualization, A.N., A.A. and A.B.; supervision, A.A.; project administration, A.A.; funding acquisition, A.B. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: AA acknowledges Russian Foundation for Basic Research for the financial support (Grant No. 20-04-01085). AdB acknowledges the French National Research Agency with grant ANR-19-CE17-0021 (BASIN) and the Indo-French Centre for the Promotion of Advanced Research / CEFIPRA for collaborative grants (numbers 5302-2).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rose, G.D. Hierarchic Organization of Domains in Globular Proteins. *Journal of Molecular Biology* **1979**, *134*, 447–470, doi:10.1016/0022-2836(79)90363-2.
2. Vriend, G.; Sander, C. Detection of Common Three-Dimensional Substructures in Proteins. *Proteins* **1991**, *11*, 52–58, doi:10.1002/prot.340110107.
3. Zehfus, M.H. Continuous Compact Protein Domains. *Proteins: Structure, Function, and Bioinformatics* **1987**, *2*, 90–110, doi:10.1002/prot.340020204.
4. Gelly, J.-C.; de Brevern, A.G.; Hazout, S. "Protein Peeling": An Approach for Splitting a 3D Protein Structure into Compact Fragments. *Bioinformatics* **2006**, *22*, 129–133, doi:10.1093/bioinformatics/bti773.
5. Berezovsky, I.N.; Tumanyan, V.G.; Esipova, N.G. Representation of Amino Acid Sequences in Terms of Interaction Energy in Protein Globules. *FEBS Letters* **1997**, *418*, 43–46, doi:10.1016/S0014-5793(97)01346-X.
6. Dułak, D.; Gadzała, M.; Banach, M.; Ptak, M.; Wiśniowski, Z.; Konieczny, L.; Roterman, I. Filamentous Aggregates of Tau Proteins Fulfil Standard Amyloid Criteria Provided by the Fuzzy Oil Drop (FOD) Model. *International Journal of Molecular Sciences* **2018**, *19*, 2910, doi:10.3390/ijms19102910.
7. Galzitskaya, O.V.; Ivankov, D.N.; Finkelstein, A.V. Folding Nuclei in Proteins. *Molecular Biology* **2001**, *35*, 605–613, doi:10.1023/A:1010535329257.
8. Lesk, A.M.; Rose, G.D. Folding Units in Globular Proteins. *Proc Natl Acad Sci U S A* **1981**, *78*, 4304–4308.
9. Wetlaufer, D.B. Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proc Natl Acad Sci U S A* **1973**, *70*, 697–701.
10. Zaki, M.J.; Nadimpally, V.; Bardhan, D.; Bystroff, C. Predicting Protein Folding Pathways. *Bioinformatics* **2004**, *20 Suppl 1*, i386-393, doi:10.1093/bioinformatics/bth935.
11. Brylinski, M.; Konieczny, L.; Kononowicz, A.; Roterman, I. Conservative Secondary Structure Motifs Already Present in Early-Stage Folding (in Silico) as Found in Serpines Family. *J. Theor. Biol.* **2008**, *251*, 275–285, doi:10.1016/j.jtbi.2007.10.041.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2021 doi:10.20944/preprints202106.0661.v1

12. Bateman, A.; Coin, L.; Durbin, R.; Finn, R.D.; Hollich, V.; Griffiths-Jones, S.; Khanna, A.; Marshall, M.; Moxon, S.; Sonnhammer, E.L.L.; et al. The Pfam Protein Families Database. *Nucleic Acids Res* **2004**, *32*, D138–D141, doi:10.1093/nar/gkh121.
13. Sheridan, R.P.; Venkataraghavan, R. A Systematic Search for Protein Signature Sequences. *Proteins: Structure, Function, and Bioinformatics* **1992**, *14*, 16–28, doi:10.1002/prot.340140105.
14. Exarchos, T.P.; Papaloukas, C.; Lampros, C.; Fotiadis, D.I. Mining Sequential Patterns for Protein Fold Recognition. *J Biomed Inform* **2008**, *41*, 165–179, doi:10.1016/j.jbi.2007.05.004.
15. Attwood, T.K.; Coletta, A.; Muirhead, G.; Pavlopoulou, A.; Philippou, P.B.; Popov, I.; Romá-Mateo, C.; Theodosiou, A.; Mitchell, A.L. The PRINTS Database: A Fine-Grained Protein Sequence Annotation and Analysis Resource—Its Status in 2012. *Database (Oxford)* **2012**, *2012*, doi:10.1093/database/bas019.
16. Via, A.; Gherardini, P.F.; Ferraro, E.; Ausiello, G.; Scalia Tomba, G.; Helmer-Citterich, M. False Occurrences of Functional Motifs in Protein Sequences Highlight Evolutionary Constraints. *BMC Bioinformatics* **2007**, *8*, 68, doi:10.1186/1471-2105-8-68.
17. Domingues, F.S.; Lengauer, T. Protein Function from Sequence and Structure Data. *Appl. Bioinformatics* **2003**, *2*, 3–12.
18. Pei, J.; Cai, W.; Kinch, L.N.; Grishin, N.V. Prediction of Functional Specificity Determinants from Protein Sequences Using Log-Likelihood Ratios. *Bioinformatics* **2006**, *22*, 164–171, doi:10.1093/bioinformatics/bti766.
19. Weiss, O.; Jiménez-Montaño, M.A.; Herzel, H. Information Content of Protein Sequences. *J. Theor. Biol.* **2000**, *206*, 379–386, doi:10.1006/jtbi.2000.2138.
20. Ptitsyn, O.B.; Volkenstein, M.V. Protein Structure and Neutral Theory of Evolution. *J. Biomol. Struct. Dyn.* **1986**, *4*, 137–156, doi:10.1080/07391102.1986.10507651.
21. Szoniec, G.; Ogorzalek, M.J. Entropy of Never Born Protein Sequences. *Springerplus* **2013**, *2*, doi:10.1186/2193-1801-2-200.
22. Yoon, B.-J. Hidden Markov Models and Their Applications in Biological Sequence Analysis. *Curr Genomics* **2009**, *10*, 402–415, doi:10.2174/138920209789177575.
23. Nekrasov, A.N. Entropy of Protein Sequences: An Integral Approach. *Journal of Biomolecular Structure and Dynamics* **2002**, *20*, 87–92, doi:10.1080/07391102.2002.10506825.
24. Nekrasov, A.N.; Anashkina, A.A.; Zinchenko, A.A. *A New Paradigm of Protein Structural Organization*; Institute of Physics, Belgrade, 2014;
25. Anashkina, A.A.; Nekrasov, A.N. The Method for Identification of Hierarchical Organization of Protein Sequences. *Russian Journal of Numerical Analysis and Mathematical Modelling* **2014**, *29*, 265–273, doi:10.1515/rnam-2014-0021.
26. Nekrasov, A.N.; Zinchenko, A.A. Hydrolases: The Correlation Between Informational Structure and the Catalytic Sites Organization. *Journal of Biomolecular Structure and Dynamics* **2008**, *25*, 553–561, doi:10.1080/07391102.2008.10507202.
27. Chertkova, R.V.; Brazhe, N.A.; Bryantseva, T.V.; Nekrasov, A.N.; Dolgikh, D.A.; Yusipovich, A.I.; Sosnovtseva, O.; Maksimov, G.V.; Rubin, A.B.; Kirpichnikov, M.P. New Insight into the Mechanism of Mitochondrial Cytochrome c Function. *PLoS One* **2017**, *12*, doi:10.1371/journal.pone.0178280.
28. Shingarova, L.N.; Petrovskaya, L.E.; Nekrasov, A.N.; Kryukova, E.A.; Boldyreva, E.F.; Yakimov, S.A.; Guryanova, S.V.; Dolgih, D.A.; Kirpichnikov, M.P. Expression and Properties of Human TNF Peptide Fragments. *Russ J Bioorg Chem* **2010**, *36*, 301–309, doi:10.1134/S1068162010030040.
29. Shannon, C.E. A Mathematical Theory of Communication. *Bell System Technical Journal* **1948**, *27*, 379–423, doi:10.1002/j.1538-7305.1948.tb01338.x.
30. Nekrasov, A.N.; Anashkina, A.A.; Zinchenko, A.A. A New Paradigm of Protein Structural Organization. In Proceedings of the Proc. 2nd Int. Conf."Theoretical Approaches to Bioinformation Systems"(TABIS 2013); Institute of Physics, Belgrade, 2014; pp. 1–22.
31. Nekrasov, A.N.; Radchenko, V.V.; Shuvaeva, T.M.; Novoselov, V.I.; Fesenko, E.E.; Lipkin, V.M. The Novel Approach to the Protein Design: Active Truncated Forms of Human 1-CYS Peroxiredoxin. *Journal of Biomolecular Structure and Dynamics* **2007**, *24*, 455–462, doi:10.1080/07391102.2007.10507133.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 June 2021 doi:10.20944/preprints202106.0661.v1

32. Briers, Y.; Miroshnikov, K.; Chertkov, O.; Nekrasov, A.; Mesyanzhinov, V.; Volckaert, G.; Lavigne, R. The Structural Peptidoglycan Hydrolase Gp181 of Bacteriophage PhiKZ. *Biochem. Biophys. Res. Commun.* **2008**, *374*, 747–751, doi:10.1016/j.bbrc.2008.07.102.
33. Nekrasov, A.N.; Petrovskaya, L.E.; Toporova, V.A.; Kryukova, E.A.; Rodina, A.V.; Moskaleva, E.Y.; Kirpichnikov, M.P. Design of a Novel Interleukin-13 Antagonist from Analysis of Informational Structure. *Biochemistry Mosc.* **2009**, *74*, 399–405.
34. Mikhailova, A.G.; Nekrasov, A.N.; Zinchenko, A.A.; Rakitina, T.V.; Korzhenevsky, D.A.; Lipkin, A.V.; Razguljaeva, O.A.; Ovchinnikova, M.V.; Gorlenko, V.A.; Rumsh, L.D. Truncated Variants of Serratia Proteamaculans Oligopeptidase B Having Different Activities. *Biochemistry Moscow* **2015**, *80*, 1331–1343, doi:10.1134/S0006297915100156.
35. de Brevern, A.G. Protein Local Conformations at the Light of a Structural Alphabet. *Biophysical Journal* **2018**, *114*, 231a, doi:10.1016/j.bpj.2017.11.1286.
36. Nekrasov, A.N.; Alekseeva, L.G.; Pogosyan, R.A.; Dolgikh, D.A.; Kirpichnikov, M.P.; de Brevern, A.G.; Anashkina, A.A. A Minimum Set of Stable Blocks for Rational Design of Polypeptide Chains. *Biochimie* **2019**, *160*, 88–92, doi:10.1016/j.biochi.2019.02.006.
37. Jurkowski, W.; Brylinski, M.; Konieczny, L.; Wiíniowski, Z.; Roterman, I. Conformational Subspace in Simulation of Early-Stage Protein Folding. *Proteins* **2004**, *55*, 115–127, doi:10.1002/prot.20002.
38. Jurkowski, W.; Kułaga, T.; Roterman, I. Geometric Parameters Defining the Structure of Proteins--Relation to Early-Stage Folding Step. *J. Biomol. Struct. Dyn.* **2011**, *29*, 79–104, doi:10.1080/07391102.2011.10507376.
39. Kaushik, R.; Zhang, K.Y.J. A Protein Sequence Fitness Function for Identifying Natural and Nonnatural Proteins. *Proteins: Structure, Function, and Bioinformatics* **2020**, *88*, 1271–1284, doi:10.1002/prot.25900.