Keldysh Institute • Publication search

Keldysh Institute preprints • Preprint No. 40, 2019

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Tikhonov D.A., Kulikova L.I.,
Efimov A.V.

The study of interhelical distances of helical pairs in protein molecules

Recommended form of bibliographic references: Tikhonov D.A., Kulikova L.I., Efimov A.V. The study of interhelical distances of helical pairs in protein molecules // Keldysh Institute Preprints. 2019. No. 40. 21 p. doi:10.20948/prepr-2019-40-e
URL: http://library.keldysh.ru/preprint.asp?id=2019-40&lg=e

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

D.A. Tikhonov, L.I. Kulikova, A.V. Efimov

The study of interhelical distances of helical pairs in protein molecules

Москва — 2019

Тихонов Д.А., Куликова Л.И, Ефимов А.В.

Исследование межспиральных расстояний в спиральных парах белковых молекул

В данной работе проведен анализ распределения межспиральных расстояний в парах связанных между собой перетяжками спиралей в пространственных структурах белковых молекул. Были разработаны правила отбора спиральных пар в структурах белковых молекул из Банка белковых структур (Protein Data Bank). Полученное множество спиральных пар было проанализировано с целью его классификации и установления закономерностей структурной организации. Предложена точечная модель двухспирального мотива. По критерию пересечения проекций спиралей на параллельные плоскости, проходящие через оси спиралей, полученное множество разбито на три подмножества. В работе получены гистограммы распределения всех типов спиральных пар в зависимости от расстояний между спиралями. Представлены статистические оценки распределений межплоскостного и минимального расстояний для спиральных пар различных типов, принадлежащих различным множествам.

Ключевые слова: структурные мотивы белков, спиральные пары в белковых молекулах, точечная модель, статистический анализ, межплоскостное расстояние, минимальное расстояние между осями спиралей.

Dmitry A. Tikhonov, Liudmila I. Kulikova, Alexander V. Efimov

The study of interhelical distances of helical pairs in protein molecules

In this paper, the study of interhelical distances in pairs of connected α-helices found in known proteins has been performed. A number of rules for selection of the helical pairs from a set of protein structures obtained from the Protein Data Bank (PDB) have been developed. The set of helical pairs has been analyzed for the purpose of classification and finding out the features of protein structural organization. A point model of a double-helix motif has been proposed. All pairs of connected helices were divided into three subsets according to the criterion of crossing of projections of the helices on parallel planes, which pass through the axes of the helices. In this work histograms of the distribution of all types of helical pairs are obtained depending on the interhelical distances. The statistical estimates of the interplanar and minimal distance distributions for helical pairs of various types belonging to different sets are presented.

Key words: structural motives of proteins, helical pairs in proteins, point model, statistical analysis, interplane distance, minimum distance between helical axes.

The study was made with the support from the RFBR (project № 18-07-01031-a).

Introduction

Currently, structural motives having a unique spatial packing of the polypeptide chain are a central preoccupation of researchers. The interest in the structural motives is due to uniqueness of their structure and their capacity to serve as nuclei to which the rest fragments of the chain can attach themselves in the course of folding in accordance with certain rules [1]. However, irrespective of the mechanism by which the protein folding occurs, the structural motives can be used to create a system of structural classification of proteins [2]. Besides they can serve as starting structures in the course of identification of possible arrangements of a polypeptide chain in modeling the protein structure [1, 3]. Usually (though not always), a structural motif is a combination of several elements of the secondary structure. The simplest structural motives consist of two elements of the secondary structure which have a unique spatial arrangement. The object of our interest is the structural motives formed of two helices arranged one after another and bound by connections [4–6].

It is well known that α-helices are packed most densely [6]. Just α-helices prevail in proteins, probably because they are the most stable element of the secondary structure. π-helices are rather few. 3_{10}-helices, mainly right-handed ones, occur as short fragments [7]. α-helix is a well-studied secondary structure [8–12] which, together with β-structure, in many respects determines their common configuration. The stabilizing elements for α-helix are alanine, leucine and methionine (Ala, Leu and Met), while glycine (Gly) deteriorates the helix and facilitates the formation of irregular fragments. Proline (Pro) is not found in α-helix either (except for its N-terminal coil). The influence of aminoacid composition on the secondary structure can be estimated not only statistically [13], or experimentally [14], but also theoretically [7]. Besides, stability of α-helix is achieved by a certain order of alternation of hydrophobic groups in a chain which results in the formation of an entirely hydrophobic surface on the helix. The importance of appropriate alternation of side groups for the formation of a protein secondary structure was shown in papers [15, 16].

Similar descriptions can be made for even more complicated structures which consist of two or three helices or a combination of several elements of the secondary structure [17].

As is known, the most compact packing of two α-helices is reached in the case of antiparallel, perpendicular and, the so-called, slanted arrangements of the helices. Examples of such packing are super-secondary structures: α-α-corners, α-α-hairpins, L-shaped and V-shaped structures [3].

It is known that α-α-corner is a frequent structural motif in proteins [6]. This super-secondary structure is formed by two neighboring α-helices bound by connections and packed orthogonally. In proteins α-α-corners occur in the form of a left-handed super-helix. Their sequences are arranged in a special way in a chain of hydrophobic, hydrophilic and glycine residues.

Super-secondary structures of two neighboring α-helices which are bound by a connection and packed in antiparallel are α-α-hairpins. This super-secondary structure can be left-handed or right-handed depending on whether the second α-helix occurs to the left or to the right of the first one. The lengths of connections between the helices can also be different. Besides each standard α-α-hairpin should have a specific and unique position in a chain of hydrophobic, hydrophilic and glycine residues [3].

L-shaped structures are also formed of two helices. A special role in their formation belongs to proline (Pro) which facilitates a break of a connection between two α-helices. These structures can be right-handed or left-handed [5].

V-shaped structures also consist of two α-helices. They look a lot like α-α-hairpins, in which unbound ends of α-helices are widely spaced; they also resemble L-shaped structures. In V-shaped structures, the length of α-helices does not usually exceed three or four coils [3].

In our earlier works [18, 19] we solved the problems of recognition [20, 21], analysis of stability and conformational analysis of structural motives of proteins of α-α-corner type in a computational experiment of molecular dynamics. The object of investigation was α-α-corners with a short connection. Earlier (in 1993) the stability of α-α-corners was indirectly proved *in vitro* [22]. The hypothesis (which we proposed independently) of autonomous stability of structural motives was checked *in silico* in computational experiments of molecular dynamics [18].

In this paper we set up two main problems:

– select from the Protein Data Bank (PDB) all the structural motives which are formed of two helices of any type, arranged one after another in a polypeptide chain, and bound by connections of different lengths and having different conformations. These structural motives will form a data base for further analysis of such two-helical structures;

– perform a statistical analysis of the distribution of interhelical distances in the helical pairs from this data base.

Point model of a helical pair

It is well known that helices are frequent in proteins [6]. There exist proteins which consist of only helices and irregular fragments of different conformations (see histogram in fig. 1). It is evident from the histogram that in the PDB there are many proteins which contain 60–70 % of helices. At the same time the histogram demonstrates that the distribution of the number of proteins depending on the ratio between the number of aminoacids in the helices and the total number of aminoacids in the structures has a maximum for 40 %.

A helical pair is a fragment of a protein chain which consists of two neighboring helices between which there is one or several aminoacids whose secondary structure is not helical. This element is probably the simplest element of the super-secondary structure. Such an object is very convenient to analyze, since it can be described by as

few as four points of space. Indeed, if we approximate both the helices by cylinders around which the helices formed by a strand passing through C_α-atoms are wound, then the original and terminal points of the cylinder axes will be the four points which fully describe this super-secondary structure. Figure 2 presents an example of a helical pair. This is a fragment of a chain from the PDB (PDB ID 3A0B, C_α: 1000–1037). The figure illustrates the cylinders approximating the helix and the planes passing through the cylinder axes. The curve is approximated by the positions of C_α-atoms of the protein chain; the atoms on the curve are shown by points.

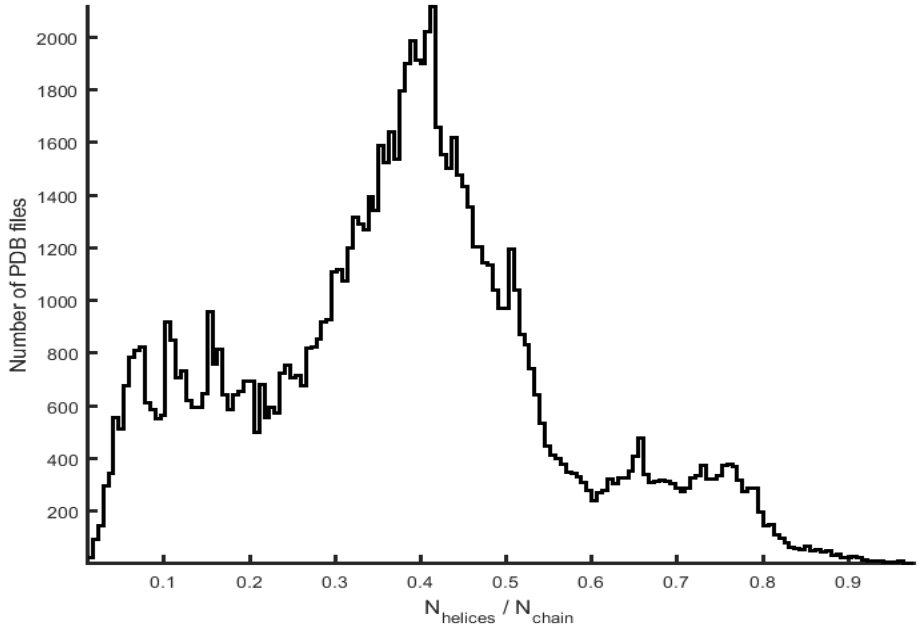

Fig. 1. Histogram of the distribution of the number of proteins in the PDB depending on the ratio between the number of aminoacids in the helices and the total number of aminoacids in a protein.

From the viewpoint of mutual arrangement of helices, three distances naturally come up. The first one is the interplane distance. As is known, one can uniquely place two parallel planes onto two noncrossing right lines so that the shortest distance between the planes be the same as the distance between the lines. Assuming that the cylinder axes lie on the right lines, the shortest distance between these right lines will be referred to as the interplane distance of a helical pair and denoted as d. The second distance of a helical pair is the shortest distance between the segments which are the cylinder axes. It will be denoted as r. Obviously, the shortest distance of a helical pair r is always greater than or equal to the interplane distance d, therefore, even formally, we can introduce the value of a leg l and determine it as $l = \sqrt{r^2 - d^2}$. The leg l will be the third distance which describes a relative arrangement of the helixes in a helical pair. The point model of a helical pair is illustrated in figure 3. Only the axes of a helical pair are shown. The segment $[A_1, A_2]$ is the axis of the cylinder of the first

helix, $[B_1, B_2]$ is the axis of the cylinder of the second helix. The figure also demonstrates all possible distances d, r and l between the helices in a helical pair.

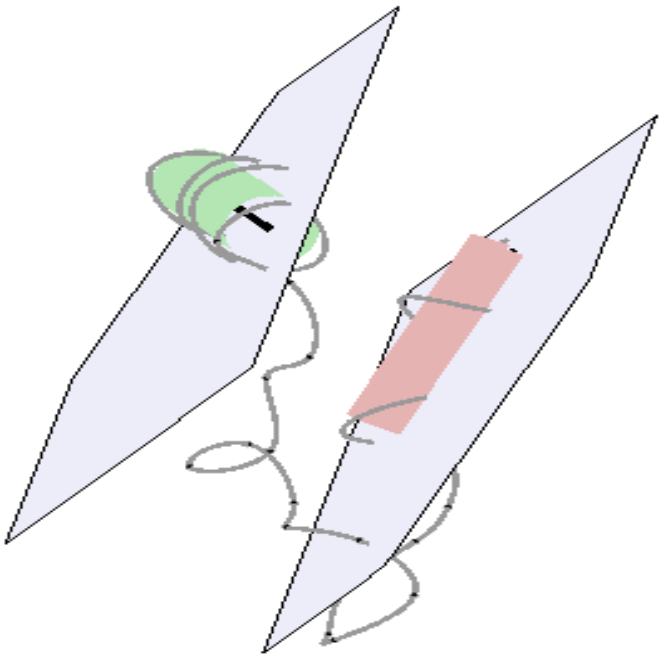

Fig. 2. Example of a helical pair. A fragment of a protein chain from the PDB consisting of 38 aminoacids (PDB ID 3A0B**,** fragment coordinates: 1000–1037). The cylinders approximating the helix and the planes passing through the cylinder axes are shown. The curve is approximated by the positions of C_α-atoms of the protein chain, the atoms on the curve are shown by points.

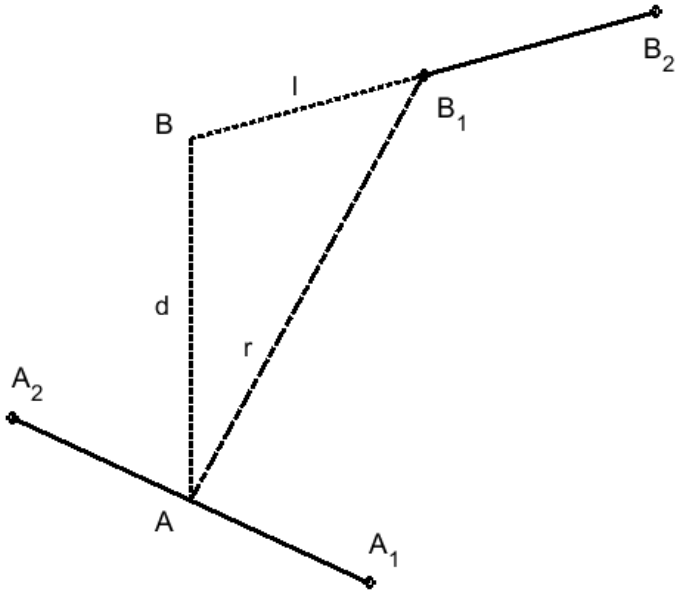

Fig. 3. Geometry of a helical pair – four points which form two segments in space.

Figure 3 illustrates a particular case of a mutual arrangement of the axes when one point from where the perpendicular joining the planes comes out (in this case this is point *A*) is an internal point of the helical axis [A_1, A_2]. The second point *B* does not belong to the helical axis [B_1, B_2]. In this case the leg *l* is a continuation of the helical axis to point *B*, where the perpendicular joining the planes comes in. Another particular case is the case when both the points *A* and *B* do not belong to their axes. Then the minimum distance between the segments is equal to the distance between the ends of the helical axes which are closest to the points from where the perpendicular comes out and where it comes in. Then the leg *l* will not have a simple geometrical representation, but can always be calculated formally.

A special case of arrangement is when both the points *A* and *B* are internal points of the axes [A_1, A_2] and [B_1, B_2] respectively. Then the interplane distance *d* coincides with the minimum distance *r*, and therefore the leg *l* is equal to zero. This is the case when the projections of the cylinder axes cross. For this case, we will consider the distribution of the interplane distance *d* in detail. When the axes cross, the ratio r/d is equal to 1. Here we investigate the general form of the statistical distribution for the ratio r/d. The statistical distributions of all the distances in a helical pair will be investigated in the same way.

Criterion of subdivision of all the helical pairs into subsets

In this work we will use the criterion of the square of the polygon of overlapping of helical axes projections onto the parallel planes passing through the helical axes in a helical pair. The set of helical pairs will be subdivided into three subsets according to the following rules:

1. Subset {*A*} involves the helical pairs not having crossing projections;
2. Subset {*B*} involves the helical pairs having crossing projections except for the helical pairs where the overlapping polygon contains the cross point of the helical axes projection;
3. Subset {*C*} involves the helical pairs for which the overlapping polygon contains the cross point of the helical axes projection.

The square of the polygon of overlapping of helical projections depends not only on geometry, but also on the cylinder diameter. Besides, the cylinder diameter is determined not only by the type of a helix but also by the value of the mean size of the side chain *DR*. The generally accepted value is *DR* = 3.6 Å. Figure 4 illustrates the polygon of overlapping of the helical projections for the helical pair shown in figure 2. The widths of the rectangles which are projections of relevant axes are equal to the diameters of the helices which are determined by their type plus double value of *DR*. In this case both the helices are α-helices with diameter 4.6 Å. Hence, the width of the rectangles is 11.8 Å. The overlapping polygon is indicated with color, its square *S* and perimeter *P* are given. The point where the projections of the helical axes cross is marked. The figure also presents the value of the interplane distance *d*.

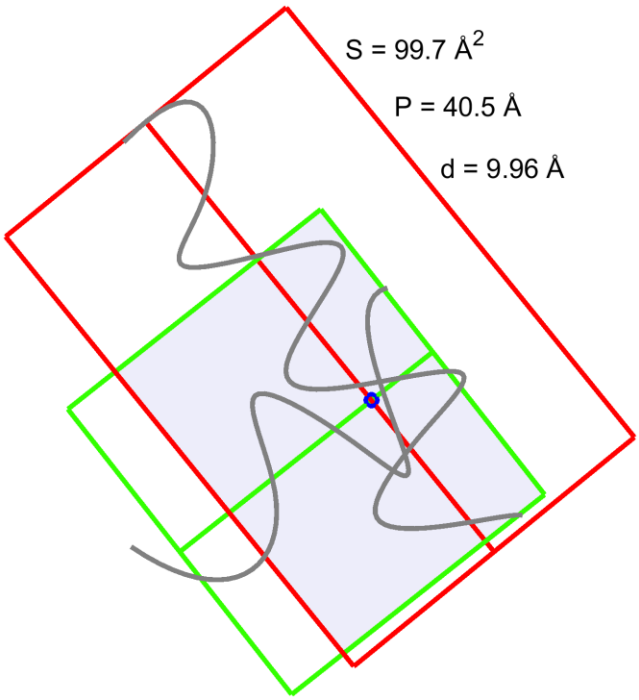

Fig. 4. Overlapping of helical cylinders projections of a helical pair. The polygon of overlapping of helical projections for the helical pair shown in figure 2 (PDB ID 3A0B**,** fragment coordinates:1000–1037).

The boundary between the subsets {*A*} and {*B*} is conditional and determined within the accuracy of *DR* value. When the value of *DR* is small, elements from the subset {*B*} will pass into the subset {*A*}. The reverse is also true. The value of *DR* has character averaged over all the aminoacid residues, it is rather approximate. Figure 4 presents the value of the interplane distance d = 9.96 Å, it is far less than 11.8 Å, that should have been for two α-helices in contact with regard to the mean size of the side chain *DR* = 3.6 Å. However, for definiteness we performed our calculations just with the use of this conventional value of *DR*.

Calculation details

For further analysis, we need point models of the helical pairs and the squares of the helical pair projections onto the parallel planes passing through the helical axes. In this section we describe how we found them.

The protein structures were taken from the Protein Data Bank (PDB). The secondary structures of animoacid chains were analyzed by the method developed by the authors of the Dictionary for Secondary Structure of Protein [23]. This method is widely used, besides the program (based on this method) which processes PDB-files and outputs a file with the secondary structure is publicly available. First of all, we were interested in the availability of such elements of the secondary structure as helices. The DSSP program distinguishes between three types of helices. The first type, designated by letter H**,** is α-helix, the second type, designated by letter G**,** is 3_{10}-helix, the third type, designated by letter I**,** is π-helix. The output files of the DSSP

program were processed with a view to find the origins and ends of continuous fragments encoded as helices. If a PDB-structure contained a protein in which there were not more than one helix, it was discarded. The structures formed of two neighboring helices of different types bound by a zero-length connection were discarded too. Hence, we selected only the structures formed of helical pairs. The data on the number of such structures are given in table 1.

Table 1.

Data – the number of protein molecules processed and the number of different-type helical pairs found

Number of processed protein structures from PDB	Number of processed aminoacid residues	Number of processed aminoacid chains	Number of helices of H type	Number of helices of G type	Number of helices of I type
100397	66546491	384666	1952658	750605	2908

The table suggests that the helices of H type prevail (72.16 %). Helices of G type account for 27.73 % of the total number of helical pairs. Helices of I type account for 0.1 % of the total number of helical pairs.

After all the helices in an aminoacid chain had been found, helical pairs were composed from neighboring helices. For example, if an aminoacid chain contains N helices, the number of helical pairs in it is $N - 1$. If we write h_i for a helix of any type inside a chain, then the helical pairs can be described as (h_i, h_{i+1}), where i varies from 1 to $N - 1$.

When the helical pairs are identified, we should find the helical axes. To do so we use the method which is widely used in molecular calculations. We know the type of the helix, the number of aminoacids n_h, therefore we can construct the coordinates of an ideal helix in a convenient coordinate system. An ideal helix is determined by three parameters: angular increment φ_h, linear step z_h and radius r_h. If we choose the axis of an ideal helix to lie along axis z, then the coordinates of C_α-atoms will be expressed in terms of the ideal helix parameters as:

$$x_i = r_h \cos(\varphi_h (i-1)),$$
$$y_i = r_h \sin(\varphi_h (i-1)),$$
$$z_i = z_h (i-1), \qquad i = 1 : n_h.$$

We also know the actual coordinates of C_α-atoms of the helices of a helical pair. To superimpose the actual and ideal coordinates let us use the algorithm proposed in papers [24, 25]. The algorithm is based on singular expansion of a covariance matrix of ideal and actual coordinates. It enables us to calculate the rotation matrix and the

translation vector of the ideal coordinates in such a way that they should best coincide with the actual coordinates. The quality of coincidence is estimated through the value of RMSD, which is a square root of the mean square of the coordinate's deviation. The rotation matrix and the translation vector are chosen from the condition that the value of RMSD be minimum. Table 2 lists the values of the parameters of ideal helices for their different types and the mean estimates of RMSD which we obtained when approximated actual helices by ideal ones.

Table 2.

Parameters of ideal helices of different types and statistical estimates of the RMSD value distribution obtained as a result of approximation of actual helices by ideal ones (φ_h – conformation angle of an ideal helix, z_h – step of an ideal helix, r_h – radius of an ideal helix)

parameters	H-helix	G-helix	I-helix
φ_h, degrees	100.00	120.00	81.82
z_h, Å	5.40	6.00	4.80
r_h, Å	2.30	1.90	2.80
mode RMSD	0.20	0.05	1.98
median RMSD	0.31	0.13	2.00
mean RMSD	0.40	0.20	2.10

As is seen, the radius of an ideal α-helix r_{hH} is equal to 2.3, the radius of an ideal G-helix r_{hG} is equal to 1.9. Analyzing the data of table 2 it is easy to see that for H and G types, approximation by ideal helices is highly satisfactory; the mean RMSD is less than the contemporary resolution in crystallography. However, for I type, the quality of approximation is not high.

Now let us discuss how we calculate the polygon of overlapping of helical projections onto the parallel planes passing through the helical axes. To find the helical axes we act upon an ideal axis which originates at the coordinate origin and ends at point $n_h z_h$ of z-axis by a rotation matrix and a translation vector. When both the axes are found, we project the second axis onto the plane where the first axis lies. The first axis together with the projection of the second axis uniquely determine the first plane. We build a normal to this plane and calculate a rotation matrix which unites the first plane and the plane xy. After rotation, both the helical axes represent segments on a plane. We reveal whether these segments intersect. Then we make projections of the helices onto the planes which are rectangles whose heights are equal to the axes lengths, and the widths are equal to double value of $r_h + DR$. Then we reveal whether these rectangles overlap. If they do, we find the polygon of their overlapping. This problem is solved with the use of MatGeom package [26], like the other problems discussed above. When the overlapping polygon is found, we

calculate its square and perimeter. Hence, for each helical pair, we reveal two things: first – whether there is a point of crossing axes projections; second – whether the helical projections intersect. In addition, we calculate the square and perimeter of the overlapping polygon of helical projections, if one exists. Relying on this information we assign a helical pair to one of the subsets {A}, {B} or {C} by the above rules. Table 3 lists the number of helical pairs of different types in the subsets {A}, {B} and {C}. On the whole there are 6 types of helical pairs: HH, GG, II, HG, HI, GI. If the name of a type consists of two similar letters, the helical pair of this type is formed by two helices of the same type. For example, the helical pair of HH-type consists of two H-helices. If a helical pair consists of helices of different types its name contains the latters of the types from which it consists. For example, HG is a helical pair containing one helix of H-type and the other helix of G-type. The data are symmetrized, i.e. if a helical pair consists of helices of different types we do not differentiate between the pairs where the sequence orders of the helices are different. For example, the helical pairs HG and GH belong to HG-type.

Table 3.

Number of different-type helical pairs in the subsets of helical pairs

Subsets of helical pairs	Types of helical pairs						Number of elements in the subset
	HH	HG	GG	HI	GI	II	
{*A*}	402912	441055	125766	1588	643	0	971964
{*B*}	570830	349024	45513	1677	244	1	967289
{*C*}	234000	31719	1598	26	9	0	267352
Total number of helical pairs per type	1207742	821798	172877	3291	896	1	**2206605**

Analyzing the data in the table we can notice some peculiarities. The total number of the helical pairs in the subset {*A*} is equal to the number of elements in the subset {*B*}, each contributing 44% to the total number of the helical pairs. The rest 12% are elements of the subset {*C*}. This means that most of helical pairs in aminoacid chains (56 %) have crossing projections.

As for the subdivision by the types of the helices, it should be noted that though H-helices prevail (72 %), helical pairs of HH-type account for as little as 54.7 % of the total number of helical pairs. Helical pairs involving G helices account for 45 %, less than 0.2 % are the pairs involving I helices.

One more peculiarity can be pointed out: in the subset {*A*} where the helical pairs do not have crossing projections, most of the pairs (45 %) belong to HG-type. In the subset {*B*}, HH-pairs are predominant (59 %). In the subset {*C*}, helical pairs of HH type are vastly predominant (87.5 %). It may be said that HH type helices are more prone to internal interaction, than the rest of the pairs. However, to relate crossing of the helical projections to propensity for internal interaction we should analyze the distributions of the distances between the helices in a pair. This will be done in the following section.

Histograms of the helical pairs distribution depending on the interhelical distances

In this section we will present the histograms of the distribution of different-type helical pairs depending on the distances between the helices as well as statistical estimates of different distances. A substantial amount of data obtained as a result of processing the PDB structures provides sufficient reliability of the results. For the helical pairs belonging to a particular subset and type, we get three sets of distances: minimum distance *r*, interplane distance *d* and the leg length $l = \sqrt{r^2 - d^2}$. For these sets, we construct histograms. Besides, we estimate the ratio *r*/*d*.

All the histograms presented in this section were calculated with uniform spacing along any distance, which varied depending on the distance type. Everywhere, we calculated the number of helical pairs N_{hh} which fall into this distance interval.

The only exclusion was the distribution for the ratio *r*/*d*. For this value the step was chosen in a logarithmic scale in view of prominent peculiarities of this distribution near zero. Besides, the distribution of *r*/*d* was normalized so that to estimate the probability density.

We will start with the simplest case when the helical pairs have crossing axes projections onto the plane. These helical pairs belong to the subset {*C*}. Figures 2 and 4 illustrate a helical pair of just this type.

Figure 5 demonstrates a histogram of the distribution of the helical pairs belonging to the subset {*C*} depending on the interplane distance *d* (in this case the interplane distance *d* is equal to the minimum distance *r*). This is the only histogram that describes the helical pairs from this subset, because the leg is equal to zero, and, accordingly, the ratio *r*/*d* = 1.

It is evident from figure 5 that the histogram has a pronounced maximum in the neighborhood of 10 Å and a subordinate maximum in the region of smaller distances for HH helical pairs. Up to 7 Å there are no helical pairs which are due to sterical limitations for the helical pairs from this subset. For the helices of all the types, the histograms have non-zero values up to 30–35 Å. The helical pairs of HG and GG types do not have a pronounced peak in the vicinity of 10 Å. They are multimode distributions. It is difficult to correlate the ordinates of their local peaks with the helical radii. Indeed, bearing in mind that the sum of the radii of HG helices is equal to 4.2 Å, we should have expected that the maximum of the histogram for HG helical

pairs would shift towards smaller values as compared to HH pairs where the sum of the radii is greater by 0.4 Å. However, this is not the case. For helices of GG type, the picture is complicated by lack of statistics, since as was mentioned above, helices of this type account for as little as nearly 0.6 % of the total number of the helical pairs belonging to the subset {*C*}.

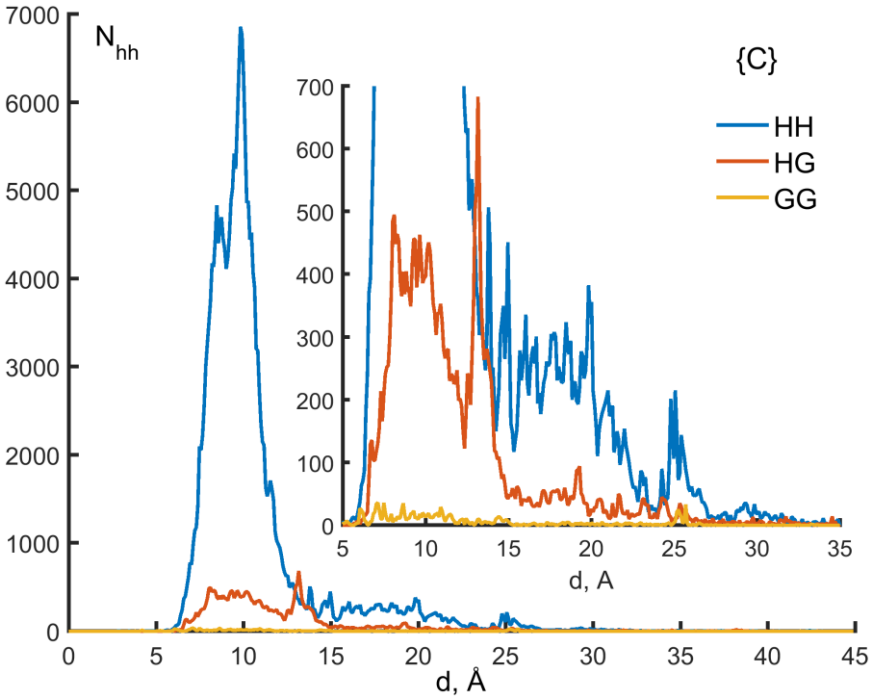

Fig. 5. Histograms of the distribution of the helical pairs belonging to the subset {*C*}, depending on the interplane distance *d* (in this case the interplane distance *d* is equal to the minimum one *r*).

Table 4 lists statistical estimates of *d* and *r* distributions for different-type helical pairs belonging to different subsets. Analyzing the data on the subset {*C*}, we can easily see that the modes (maxima) of the distributions and the mean values of the distances do not demonstrate the qualitative behavior that could be expected relying on the sum of the radii of different-type helices. However, it should be noted that for all the three types of helical pairs the distributions are sufficiently well localized since the root-mean-square deviation from the mean distances is less than half the mean distances *per se*.

Using the data of table 4 on the mean values of the distances we can try to re-estimate the conventional value of the mean size of side groups *DR*. The quantity *DR* can be related with the mean distance by the simple formulae:

$$DR_{HH} = d\_HH/2 - r_{hH},$$
$$DR_{HG} = d\_HG/2 - r_{hH}/2 - r_{hG}/2,$$
$$DR_{GG} = d\_GG/2 - r_{hG}.$$

When we substitute the values from tables 2 and 4 into these formulae we will get three *DR* values for helical pairs of different types: DR_{HH} = 3.04 Å, DR_{HG} = 3.95 Å

and DR_{GG} = 4.35 Å. The conventional value of DR_{HH} = 3.6 Å turns out to be overestimated for HH pairs and underestimated for HG and GG pairs.

Table 4.

Statistical estimates of the distributions of interplane *d* and minimum *r* distances for helical pairs of different types belonging to different subsets

Statistical estimates	*d*_HH	*d*_HG	*d*_GG	*r*_HH	*r*_HG	*r*_GG
mode {*A*}	0.88	0.41	0.41	11.88	10.56	8.86
median {*A*}	5.55	6.71	8.14	14.78	15.72	19.20
mean {*A*}	7.68	8.74	10.42	15.79	17.46	20.39
rms deviation {*A*}	7.32	7.83	8.53	8.35	9.15	9.80
mode {*B*}	8.10	0.23	4.40	9.20	8.47	5.11
median {*B*}	7.76	6.64	5.63	9.20	8.66	8.23
mean {*B*}	8.02	7.57	7.98	10.07	9.80	10.17
rms deviation {*B*}	5.71	6.13	7.22	4.77	5.12	6.26
mode {*C*}	9.82	13.20	25.68	9.82	13.20	25.68
median {*C*}	9.78	10.83	10.34	9.78	10.83	10.34
mean {*C*}	10.68	12.11	12.49	10.68	12.11	12.49
rms deviation {*C*}	3.66	4.84	6.59	3.66	4.84	6.59

Figure 6 illustrates histograms of the distribution of helical pairs of all the types belonging to the subsets {*A*} and {*B*} depending on the interplane (left column) and minimum (right column) distances. It is evident that for the helical pairs from {*A*}, the curve of the distribution as a function of the interplane distance is a monotone function which resembles the exponential one. This is the case for the helical pairs of all the types. For the helical pairs from the subset {*B*}, the function has a local maximum in the vicinity of 10 Å. Notice, that for HH-pairs, this maximum is the maximum for the whole distribution. For the other types, this maximum is local. Thus, for GH pairs, the local maximum is the vicinity of 10 Å is preserved, for the pairs of GG type, it is not. As for the long-range behavior of the interplane distance

distribution, it is approximately equal for the pairs from {*A*} and {*B*}, in both the cases the distribution extends to the value of 35–40 Å.

The right column in figure 6 is the distribution of the helical pairs belonging to the subsets {*A*} and {*B*} depending on the minimum distance between the helical axes. Here we also observe pronounced differences in the behavior of the distributions. As distinct from the left column, the distribution curve starts not at zero, but at a certain value which is determined by a minimum radius to which the axes of two helices can approach one another. It is approximately equal to 2 Å in both the cases and is related to the radius of van der Waals interaction of the atoms in the helices. The distribution of the helical pairs from the subset {*A*} depending on the minimum distance is rather a wide asymmetric distribution resembling gamma-distribution. The long-range behavior is considerable; the distribution extends beyond 45–50 Å.

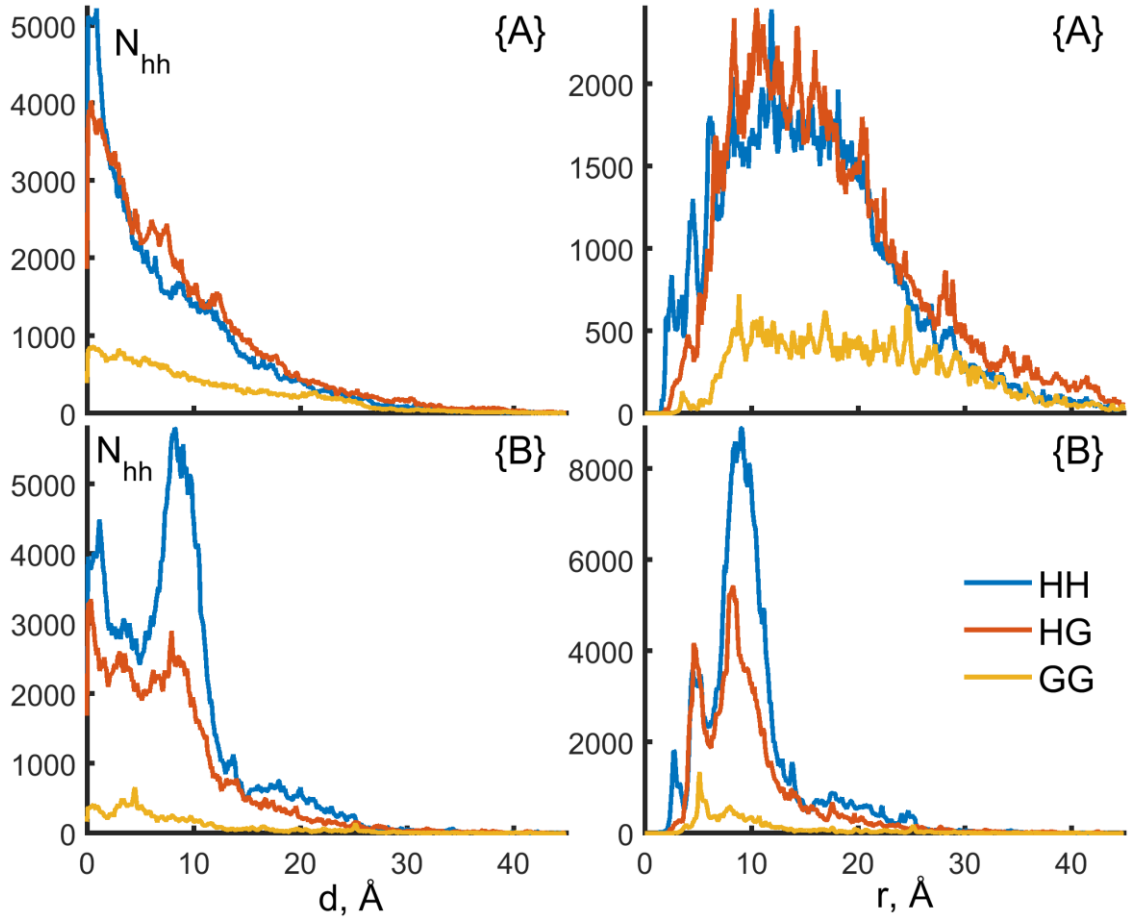

Fig. 6. Histograms of the distribution of helical pairs of different types belonging to the subsets {*A*} and {*B*} depending on the interplane distance (left column) and minimum distance (right column).

The main parameters of the histograms shown in figure 6 are presented in table 4. The data of the table, on the whole, confirm the conclusions made in comparing the distribution histograms. It can only be noted that the estimates of the mean values of the interplane distances are close in value to the dispersion of the helical pairs from the subset {*A*}, at the same time, for the minimum distances, the dispersion, on the

average, is half the mean. For the pairs from the subset {*B*}, the dispersion of the minimum distances as well as the the interplane ones is always less than the mean values of the distances. As for comparison of the mean minimum distances, for the pairs from {*A*}, they are 1.5 to 2 times greater than for the helical pairs belonging to the subset {*B*}.

Coming back to comparison of the histograms in figure 6, we can make one important conclusion: there is a qualitative difference between the interplane distance histograms, since for the helical pairs not having crossing projections (subset {*A*}), the distribution has monotone character, as distinct from the helical pairs having crossing projections (subset {*B*}). As for the minimum distances, we cannot see any qualitative differences between the histograms, unless some quantitative differences associated with long-range behavior.

Now let us consider derivatives from the minimum and interplane distances: the value of the leg which characterizes the distance between the helical projections on the plane, and the logarithm of the probability density for a ratio between the minimum and interplane distances.

Table 5.

Statistical estimates of the distribution of *r/d* and $l = \sqrt{r^2 - d^2}$ values for different helical pairs belonging to the subsets {*A*} and {*B*}

Statistical estimates	*r/d*_HH	*r/d*_HG	*r/d*_GG	*l*_HH	*l*_HG	*l*_GG
mode {*A*}	1.17	1.20	1.17	7.94	8.28	8.27
median {*A*}	2.36	2.19	2.04	11.43	11.92	14.41
mean {*A*}	15.32	22.38	32.36	12.68	13.88	16.06
rms deviation {*A*}	519.44	3869.94	7241.39	6.76	7.63	8.52
mode {*B*}	1.00	1.00	1.00	4.26	4.22	3.84
median {*B*}	1.16	1.22	1.23	4.36	4.51	4.51
mean {*B*}	6.72	11.02	5.75	4.51	4.64	4.67
rms deviation {*B*}	718.61	1360.74	131.78	2.62	2.39	2.24

Figure 7 (on the left) presents histograms of the distribution of different-type helical pairs from the subsets {*A*} (top figure) and {*B*} (bottom figure) depending on the leg. Comparing these two figures we observe some fundamental differences. The first one is that for the pairs from {*B*}, the leg has a lot of near-zero values, as distinct

from the pairs from {*A*} whose histogram has a zero value at the coordinate origin. The second difference is that for the pairs from {*A*}, the histogram is a wide asymmetric distribution which resembles the distribution of the pairs depending on the minimum distance for the same set (see fig. 6). On the contrary, the histogram of the distribution of the pairs from {*B*} depending on the leg is defined within a limited range of leg values, the leg maxima being different for different types of helices.

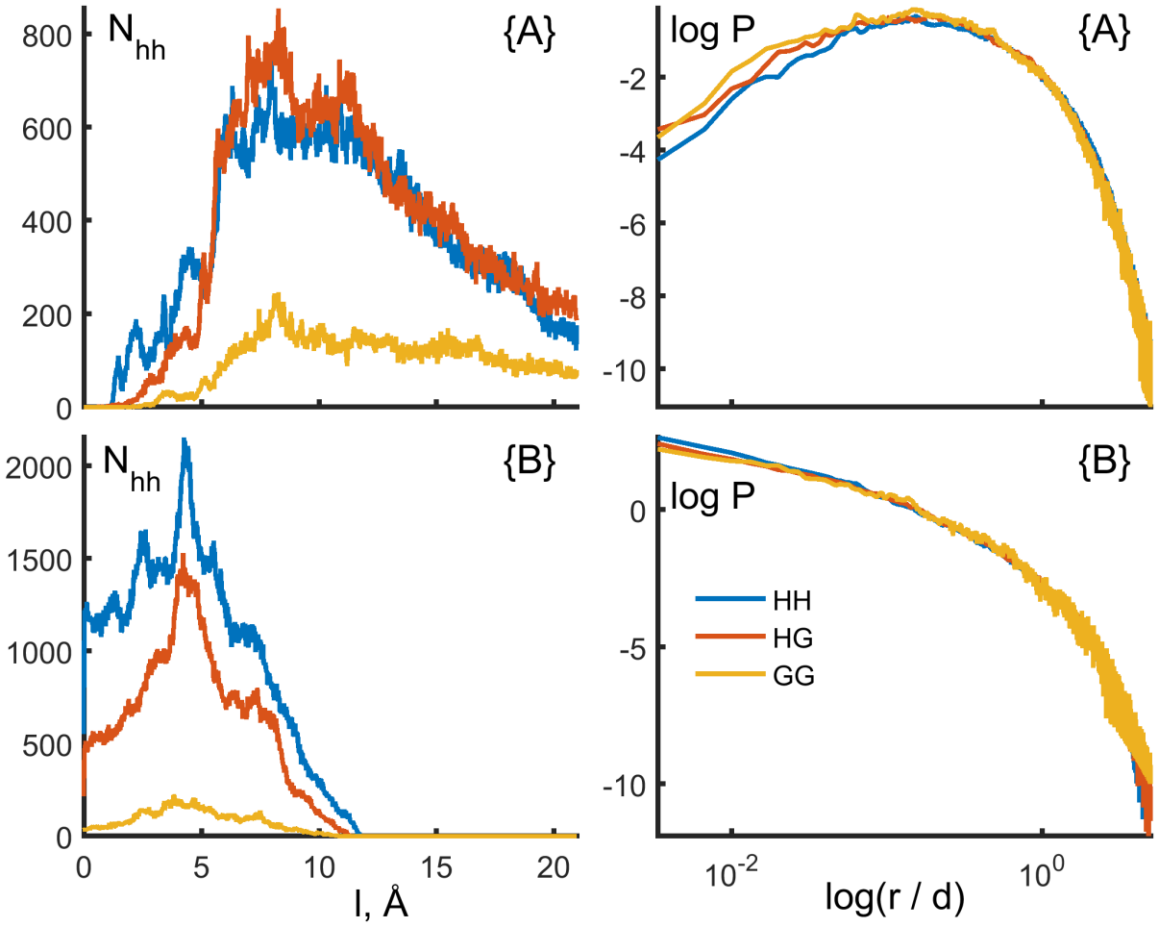

Fig. 7. Histogram of the distribution of helical pairs of all the types depending on the leg $l = \sqrt{r^2 - d^2}$ (on the left) and the logarithm of the probability of *r*/*d* ratio, calculated for each helical pair from the subsets {*A*} and {*B*} (on the right).

These maximum values are equal to 11.77 Å, 11.37 Å and 10.95 Å for HH, HG and GG types, respectively. These values are pretty close to the theoretical maximum values 11.8 Å, 11.4 Å and 11 Å, obtained on the assumption that the maximum value of the log is equal to the minimum distance to which two helices can approach one another if their axes are parallel. If the leg value turns out to be larger, a helical pair no longer has crossing projections and therefore belongs to the subset {*B*}. Table 5 lists the main parameters of the histograms presented in figure 7. Comparing the mean values and dispersions for the leg of the helical pairs from {*A*} with the mean values of the minimum distance for the same subset we can conclude that the leg

distribution demonstrates a more short-range behavior. This distribution resembles qualitatively the distribution of the pairs depending on the minimum distance, both the distributions have the character of gamma distribution. In what follows we will test this assumption.

Now let us analyze the algorithm of the probability density for the ratios between the minimum and interplane distance. Figure 7 (on the right) demonstrates these ratio algorithms for helical pairs of different types belonging to different subsets. The logarithms of the probability density are presented as functions of the r/d ratio logarithm. For helical pairs belonging to different subsets, the probability densities differ greatly only near zero value of $\log(r/d)$, therefore the abscissa of the graphs is chosen in a logarithmic scale. For the pairs from $\{A\}$, the fundamental peculiarity of the probability density is its nonmonotone character. For the pairs from $\{B\}$, the probability density is a monotone decreasing function. Table 5 lists the main characteristics of the distributions. Maximum of the probability density (mode value) for the pairs from $\{A\}$ is observed for the ratio $r/d = 1.17$. Since the abscissa scale is logarithmic, we cannot say anything about the asymptotics of the probability density. In what follows we will show that this asymptotics is proportional to $P(r/d) \to 1/(r/d)^2$ for all the helical pairs regardless of the subset they belong to. Just this fact does not enable us to calculate the mean, since its value logarithmically diverges as the sample size grows. Table 6 lists the estimates of the mean values for a finite sample and the estimates of their dispersion. Great values of the dispersion, as compared to the mean ones suggest the fact of divergence.

In the next section we will analyze the distance distribution histograms in greater detail, in particular we will test our assumption that the histograms of the minimum distance distributions can be described by gamma distribution formula. We will prove that this formula also describes the log distribution for the helical pairs from the subset $\{A\}$. We will analyze the distribution of the interplane distances and r/d ratios. We will propose a model of a random geometry of a helical pair which on the whole describes the distribution of r/d, especially the asymptotic behavior of this distribution. Finally, we will analyze the main peculiarities of r/d distribution for the helical pairs found experimentally with the use of a random geometry model.

Conclusions

Using a point model of helical pairs, we selected a set of protein molecule structures to be investigated. The set was selected from the Protein Data Bank with the use of special rules and subdivided into three subsets according to the criterion of crossing helix projections on the parallel planes passing through the axes of the helices. We analyzed the statistical properties of different distances between neighboring helices in protein chains.

In this work histograms of the distribution of all types of helical pairs are obtained depending on the interhelical distances. The statistical estimates of the interplanar

and minimal distance distributions for helical pairs of various types belonging to different sets are presented.

The authors wish to thank O.V. Sobolev for his assistance in getting the data.

References

1. Efimov A.V. Standard structures in proteins. *Prog. Biophys. Molec. Biol*. 1993; 60:201–239. doi: 10.1016/0079-6107(93)90015-C
2. Gordeev A.B., Kargatov A.M., Efimov A.V. PCBOST: Protein classification based on structural trees. *Biochemical and Biophysical Research Communications*. 2010; 397:470–471. doi: 10.1016/j.bbrc.2010.05.136
3. Efimov A.V. Super-secondary structures and modeling of protein folds. In: *Methods in Molecular Biology*. Ed. Kister A.E. Clifton: Humana Press, 2013. V. 932. P. 177–189.
4. Brazhnikov E.V., Efimov A.V. Structure of α-α-hairpins with short connections in globular proteins. *Molecular Biology*. 2001;35(1):89–97. doi: 10.1023/A:1004859003221
5. Efimov A.V. L-shaped structure from two α-helices with a proline residue between them. *Mol. Biol. (Moscow*). 1992;26:1370–1376 (in Russ.).
6. Efimov A.V. A new super-secondary protein structure: the α-α-angle. *Mol. Biol*. (*Moscow*). 1984; 18:1524–1537 (in Russ.).
7. Finkelstein A.V. *Fizika belkovykh molekul* (Physics of Protein Molecules). Moscow–Izhevsk; 2014. 424 p. (in Russ.).
8. Ptitsyn O.B., Finkelstein A.V. In: *Itogi nauki i tekhniki* (Results in Science and Technology). Ed. Vol'kenshtein M.D. Moscow; 1979. V. 15. P. 6–41. (Series “Molecular Bilogy”) (in Russ.).
9. Shul'ts G.E., Shirmer R.Kh. *Printsipy strukturnoi organizatsii belkov*. Moscow; 1982. 354 p. (Translation of: Schulz G.E., Schirmer R.H. Principles of Protein Structure Springer-Verlag New. York, 1979 (Series “Springer Advanced Texts in Chemistry”)).
10. Miller S., Janin J., Lesk A.M., Chothia C. Interior and surfage of monomeric proteins. *J. Molecular Biology*. 1987; 196:641–656. doi: 10.1016/0022-2836(87)90038-6
11. Creighton T.E. *Proteins*. 2-nd edn. N.Y.: W.H. Freeman & Co; 1991.
12. Stepanov V.M. *Molekuliarnaia biologiia. Struktura i funktsii belkov* (Molecular biology. The structure and function of proteins). Moscow; 1996. 336 p. (in Russ.).
13. Finkelstein A.V. *Molekuliarnaia biologiia* (Mol. Biol. (Moscow)). 1970;11:811–819 (in Russ.).
14. Fersht A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. N.Y.: W.H. Freeman & Co; 1999.

15. Lim V.I. Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J. Molecular Biology*. 1974; 88:857–872. doi: 10.1016/0022-2836(74)90404-5
16. Lim V.I. Algorithm for prediction of α-helices and β-structural regions in globular proteins. *J. Molecular Biology*. 1974; 88:873–894. doi: 10.1016/0022-2836(74)90405-7
17. Wierenga R.K., Terpstra P., Hol W.G.S. Prediction of the occurrence of the ADP-binding βαβ-fold in proteins, using an amino acid sequence fingerprint. *J. Molecular Biology*. 1986; 187:101–107. doi: 10.1016/0022-2836(86)90409-2
18. Rudnev V.R., Pankratov A.N., Kulikova L.I., Dedus F.F., Tikhonov D.A., Efimov A.V. Recognition and stability analysis of structural motifs of α-α-corner type in globular proteins. *Mathematical Biology and Bioinformatics*. 2013;8(2):398–406 (in Russ.). doi: 10.17537/2013.8.398
19. Rudnev V.R., Pankratov A.N., Kulikova L.I., Dedus F.F., Tikhonov D.A., Efimov A.V. Conformational analysis of structural motifs of α-α-corner in the computational experiment of molecular dynamics. *Mathematical Biology and Bioinformatics*. 2014;9(2):575–584 (in Russ.). doi: 10.17537/2014.9.575
20. Dedus FF, Kulikova LI, Pankratov AN, Tetouev RL. *Klassicheskie ortogonal'nye bazisy v zadachakh analiticheskogo opisaniia i obrabotki informatsionnykh signalov* (Classical Orthogonal Bases in Problems of Analytical Description of Information Signals and Their Processing). Moscow; 2004. 147 p. (in Russ.).
21. Pankratov A.N., Gorchakov M.A., Dedus F.F., Dolotova N.S., Kulikova L.I., Makhortykh S.A., Nazipova N.N., Novikova D.A., Olshevets M.M., Pyatkov M.I., Rudnev V.R., Tetuev R.K., Filippov V.V. Spectral Analysis for identification and visualization of repeats in genetic sequences. *Pattern Recognition and Image Analysis*. 2009;19(4):687–692. doi: 10.1134/S105466180904018X
22. Tsai F.C., Sherman J.C. Circular dichroism analysis of a synthetic peptide corresponding to the α,α-corner motif of hemoglobin. *Biochemical and Biophysical Research Communications*. 1993;196(1):435–439. doi: 10.1006/bbrc.1993.2268
23. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22(12):2577–2637. doi: 10.1002/bip.360221211
24. Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*. 1976; 32:922–923. doi: 10.1107/S0567739476001873
25. Kabsch W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*. 1978; 34:827–828. doi: 10.1107/S0567739478001680
26. Legland D. *MatGeom: Matlab geometry toolbox for 2D/3D geometric computing*. http://github.com/dlegland/matGeom (accessed 11 March 2016).
27. Holland P.W., Welsch R.E. Robust regression using iteratively reweighted least-squares. *Communications in Statistic – Theory and Methods*. 1977; 6(9):813–827. doi: 10.1080/03610927708827533

Contents