DOI 10.1007/s10958-017-3477-2
Journal of Mathematical Sciences, Vol. 225, No. 4, September, 2017

ANALYZING THE DATA BANK OF PROTEINS SPACE STRUCTURES (PDB): A GEOMETRICAL APPROACH

E. A. Vilkul, A. O. Ivanov, A. S. Mishchenko, Th. Yu. Popelenskii, A. A. Tuzhilin, and K. V. Shaytan UDC 514.8+51-76+57.087

Abstract. A geometrical approach to the analysis of a data bank containing information on space structures of proteins is suggested. As a result of the analysis, several relationships concerning possible conformations are found, and also there is obtained a list of polypeptides whose structures differ essentially from typical concepts of the 3D-structure of proteins. Those deviations could indicate possible errors and gaps of the interpreting models in use, as well as some new unknown relations and rules.

1. Introduction

In the present paper, we demonstrate the potential of a geometrical approach to verification and analysis of information on space structure of proteins. Today the study of biopolymers' space structure is one of the most important directions in the development of biology. Modern experimental methods, such as X-ray analysis [3], NMR (nuclear magnetic resonance) [14], and cryoelectron microscopy [1], give an opportunity to study and describe the space structure of many thousands of proteins. This information is collected, systematized, and stored in electronic informational banks, such as the most popular and well-known Protein Data Bank (PDB). As far back as in the 1990s, the number of protein structures represented in PDB had exceeded 5000, and by that time questions had arisen concerning finding files containing errors and/or inaccuracies, concerning possible reasons of such inaccuracies, and on the significance of such errors for the further use of those files [2, 4]. However, let us note that these problems had not got a wide and universal discussion. There are several publications (see [2, 4]) on the analysis of errors contained in PDB, but these papers just briefly mention the presence of some errors usually and do not go into details. Probably due to this reason this topic remains unnoticed by several leading specialists. As a corollary, the current version of PDB remains unsuitable for direct using. The typical situation is that before using the PDB data specialists have to apply some pre-processing procedures based on molecule dynamics technique.

We must note that each file that is supposed to be deposited in the PDB must pass a sequence of tests. However, for today PDB still contains many files with information that contradicts common concepts on the structure of proteins and amino acids. In our previous publications [5–7], we used simple geometrical models for macromolecules geometry analysis. We tested all the data from PDB obtained by NMR (about 10 000 files). As a result of this selection (probably an excessively strict one) we get a small but more reliable data bank (about 3 000 files); see details in Sec. 3. Note that a similar selection also proceeded for the X-ray analysis part of PDB (see [13]). The aim of the present paper is to study deviations from the well-know "Plane Law" for neighboring amino acids from our short NMR data bank. Recall (see details in Sec. 4) that in accordance with this law, six atoms from each pair of consecutive amino acids (three form the first one and three from the second one) are located in the same two-dimensional plane. We study deviations from this law by means of a simple geometric test based on the calculations of the volume of the convex hull of these six atoms. It turns out that even among the selected files the deviations can be quite essential. Analyzing the deviations found, we discovered that such deviations are often located

Translated from Fundamentalnaya i Prikladnaya Matematika, Vol. 20, No. 3, pp. 33–46, 2015.

at the areas between the typical elements of the secondary protein structure (the so-called helices and sheets).

In the future, it would be interesting to analyze possible relations between the deviations form the Plane Law and locations of the fragments of protein active sites.

2. The Model of a Protein Molecule

A protein molecule can be described as a space geometric graph, whose vertices correspond to the centers of atoms and whose edges, which are realized by straight segments, correspond to covalent bonds.

A protein molecule is a chain of *amino acid residues* having a special form, and the residues could be modified. Moreover, components of non-amino acid nature can be added to a chain. In the present paper, we consider neither modified amino acids nor other additions, and restrict ourselves by "blocks" each of which belongs to one of the 20 types of so-called standard amino acids. In Fig. 1, the names and structural formulas of these amino acids are shown.

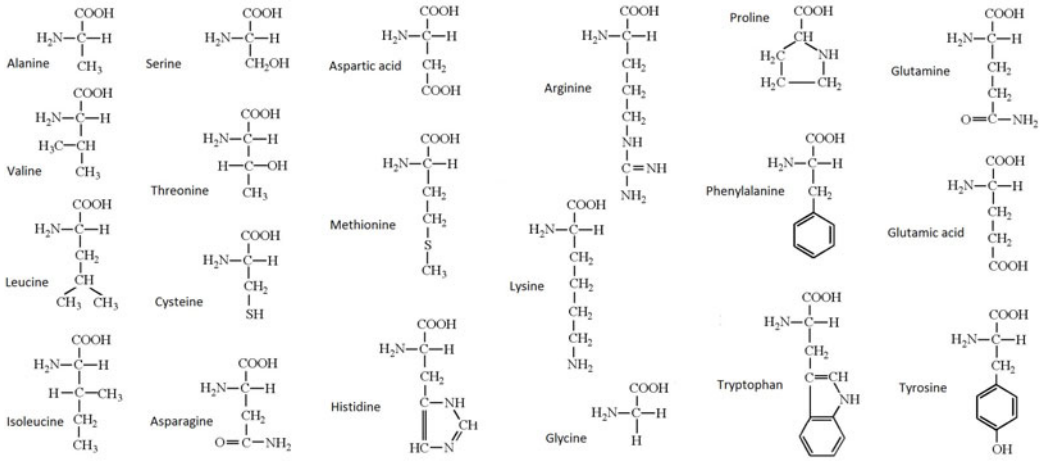

Fig. 1. The structural formulas of the 20 standard amino acids.

Each amino acid has the following structure. In its "center," a carbon atom C is located. It is usually referred to as the *alpha-carbon* (i.e., the first or the basic carbon) and is denoted by C_α or by CA; it is bonded to the *carboxyl group* COOH, *amino group* H_2N, and a carbon atom H. The remaining part of the amino acid, which is referred to as its *radical* or its *side chain*, plays the role of the "face" of the amino acid. The radical is bonded by the single covalent bond to the atom C_α (see Fig. 1). The unique exception is proline that has another bond $N - C_\delta$ that leads to a cyclic appearance. Note that, in the strict sense, proline is not an amino acid but an imino acid.

In protein formation, amino acids join to each other in a chain by the following rule: the carboxyl group of the preceding amino acid interacts with the peptide group of the posterior amino acid with the loss of a water molecule H_2O (see Fig. 2). As a result, the carbon atom from the carboxyl group bonds with the nitrogen atom from the amino group. The resulting bond is referred to as a *peptide* bond. Thus, if we cut the graph of the protein along all the peptide bonds-edges, then it splits into blocks that are the *amino acid residues*. The ending amino acid residue containing the amino group is called the N-*ending* in contrast to the second ending residue that contains the carboxyl group and is referred to as the C-*ending*. The remaining amino acid residues are called *interior*.

If one throws away all the radicals, then one obtains the *carbon–nitrogen backbone*. It can be represented as a space polygonal line of the form

$$\cdots - \mathrm{N} - \mathrm{C}_\alpha - \mathrm{C} - \mathrm{N} - \mathrm{C}_\alpha - \mathrm{C} - \cdots - \mathrm{N} - \mathrm{C}_\alpha - \mathrm{C} - \cdots,$$

with hydrogen and oxygen atoms attached to it in the corresponding places.

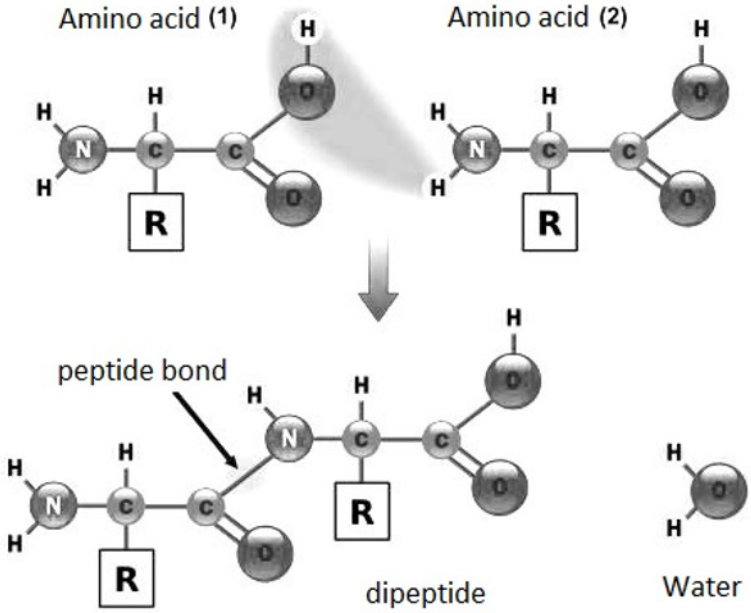

Fig. 2. Peptide chain formation.

The space structure of protein is called its *conformation*. It cannot be an arbitrary one. All the covalent bonds are considered as sufficiently rigid, i.e., their lengths cannot deviate from some special values depending on the type of the bond. The same is true for the angles between the neighboring covalent bonds that are referred to as the *covalent angles*. Further, the bonds neighboring to the peptide bond form a sufficiently flat configuration (we call this statement the *Plane Law*). At least, the atoms C_α that are incident with the peptide bond are usually located by the "different sides" of the straight line passing through the peptide bond (under the assumption that the Plane Law is valid). In this case, the amino acid residues are said to be in the *trans-configuration*. But sometimes the second possible location that is referred to as the *cis-configuration* also appears.

Thus, when studying possible conformations, the graph of a protein molecule can be represented as a linkage assembled from sufficiently elastic straight rods such that the angles between neighboring rods are sufficiently rigid. On the configuration space of this linkage an energy functional can be defined that takes into account different noncovalent interactions. The resulting conformations are determined as the ones corresponding to minima of this functional. As the dimension of the configuration space becomes big, the functional could have very many local minima (see [11, 12]). However, various experiments imply that usually there are a very small number of "deep" minima (often just a single one) corresponding to stable space structures of the protein. The problem of finding a "basic" conformation in terms of the amino acid sequence forming the peptide chain has primary value both for biology and for medicine. Its experimental solution is a very expensive, labor- and time-consuming procedure; therefore adequate computer modeling can be extremely useful here.

3. Preliminary Selection of Files

The files from the PDB contain records about all atoms of proteins, their space coordinates, and some additional information. Using the coordinates, one can easily reconstruct all the necessary details of the space structure of the protein. For example, one can calculate bond lengths, bond angles, and other geometric characteristics.

We selected files with structures obtained by NMR (it is worth noting that such files contain coordinates of the hydrogen atoms). There are approximately 10 000 files of this type. There are several different types of records describing the same amino acid. For example, for $\approx$ 59 000 glycine entries the corresponding collection of atoms was encoded as $\mathrm{N, CA, C, O, H, HA2, HA3}$, while 870 times it was encoded as $\mathrm{N, CA, C, O, H1, H2, H3, HA2, HA3}$. For this reason, we reduced the number of files in our selection by choosing files with most common encodings of the amino acids. This selection contains $\approx$ 3 000 files. This step is necessary because our geometry-based approach uses records of amino acid residues that should be standardized in some way.

After that, we perform some other tests and narrow our selection to files that have passed the tests. The tests are:

(1) covalent bonds joining two atoms in amino acids of a fixed type should have close lengths;
(2) bond angles in amino acids of a fixed type should have close values;
(3) consecutive internal amino acids are in the trans-configuration.

Estimation of length of covalent bonds was performed in the following way. In all selected files for all internal amino acids (e.g., for all glycines) we calculated mean values of lengths of covalent bonds (for all pairs of bonded atoms). Then for every peptide containing amino acid of the fixed type for covalent bonds in all occurrences of the amino acid the maximal relative deviation (in percents) from the mean value was calculated. It appears that despite all tests being performed before any structure file is deposited in the PDB, there are some files with enormously huge deviations of bond lengths from the corresponding mean values. The largest deviation 705.819% was found for the $\mathrm{N-H}$ bond in alanine (the file 2PDE.pdb [9]). In Fig. 3, one can see some "amino acids" with defective structure from this file.

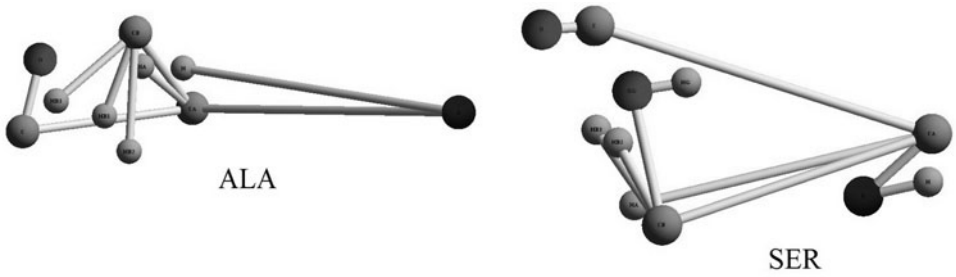

Fig. 3. Examples of defective "amino acids" from 2PDE.pdb.

Statistical analysis of bond lengths in all 20 amino acid residues for all selected 2 862 proteins showed that for most occurrences of amino acid residues deviations from the mean values do not exceed 5%. Nevertheless, for every amino acid there are files with more than 5% deviation. The number of such files varies and depends on the type of the amino acid residue (it is maximal for tryptophan — 183 files, and it is minimal for methionone — 15 files):

```
GLY: 28; ALA: 26; SER: 47; CYS: 10; PRO: 157; VAL: 37; THR: 43;
ILE: 31; LEU: 38; ASP: 29; ASN: 33; GLU: 41; GLN: 30; MET: 15;
LYS: 53; ARG: 48; HIS: 122; PHE: 26; TYR: 42; TRP: 183.
```

The number of files containing amino acid residues with more than 5% deviation is equal to 442 (among 2 862 files), which is more than 15% of all selected files. In Fig. 4, the distributions of maximal deviations are shown for all bonds, for the $\mathrm{C_\alpha - C}$ bond, and for the $\mathrm{C-N}$ bond.

The same approach was taken for bond angles. After performing all the tests, we obtained approximately 2 500 files for which the plane law is planned to be checked.

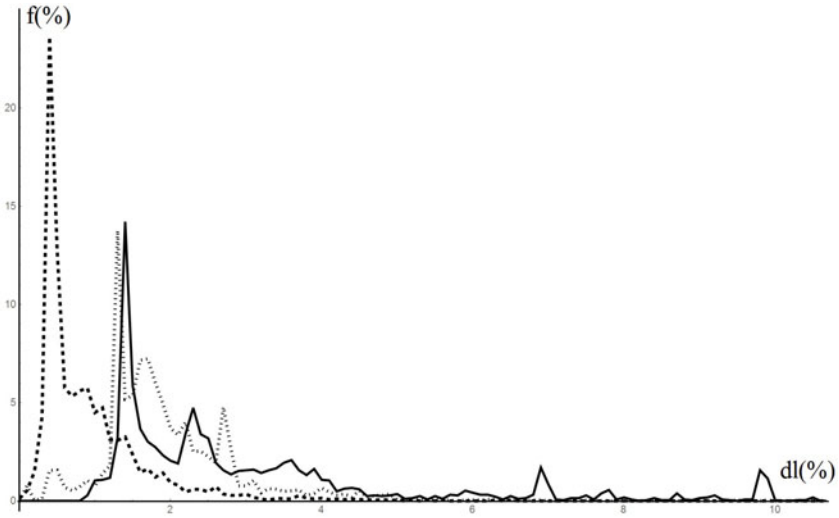

Fig. 4. The number of files (in percent, vertical axis) vs. deviation from the mean value (in percent, horizontal axis) — dashed line for all bonds, solid line for the $C_\alpha - C$ bond, and dotted line for the $C - N$ bond.

4. The Plane Law

Recall that the plane law states that six atoms, namely the atoms C_α, C, O of one amino acid residue and the atoms C_α, N, H of the next residue for most files belong to a 2-dimensional plane [10]. In the case where the next residue is proline, instead of the hydrogen atom one should consider C_δ.

This statement was checked in the following way. For every pair of consecutive residues the convex hull of peptide group was constructed and its volume was calculated. Obviously for almost plane peptide groups the volume should be close to zero. Our experiments with computer show that for volumes exceeding $1A^3$ the corresponding peptide group is significantly far from being plane. The volumes of such convex hulls were calculated for all pairs of consecutive internal amino acid residues for NMR-files from the PDB. For most peptide groups the plane law is valid. But nevertheless there are sufficiently large deviations. In Fig. 5 maximal values of the volume for all consecutive pairs of amino acid residues are shown.

Let us note that considering all the NMR-files from the PBD presumably demands some special attention due to the possible correlation of the convex hulls volumes for the peptide groups with deviations from the mean values of the bond length and the bond angles.

Further, for the selected 2 836 files the analysis shows that peptide groups with the convex hull volume exceeding $1A^3$ are found in 475 files (which is 16.7% of 2 836). On the other hand, the selected 2 836 files contain 190 185 consecutive pairs of internal amino acid residues, while the number of pairs for which the volume of a peptide group's convex hull is less than $0.575A^3$ is 95% of their total number. Hence the plane law is strong enough. But let us stress that anomalies are not so rare. The corresponding maximal volumes are shown on Fig. 6. The difference in Figs. 5 and 6 clearly shows that for our selection of the NMR-files from the PDB, the deviation from the plane law reduces significantly.

5. Regularity in Deviations from the Plane Law

The approach to the plane law checking that was described above needs a slight correction in the case where the second residue is proline. The problem is that if at the second place one has proline, then in the corresponding atom triple instead of the hydrogen atom one uses the atom C_δ, and this has significant impact on results because of different mean values of $N - H$ and $N - C_\delta$ covalent bonds.

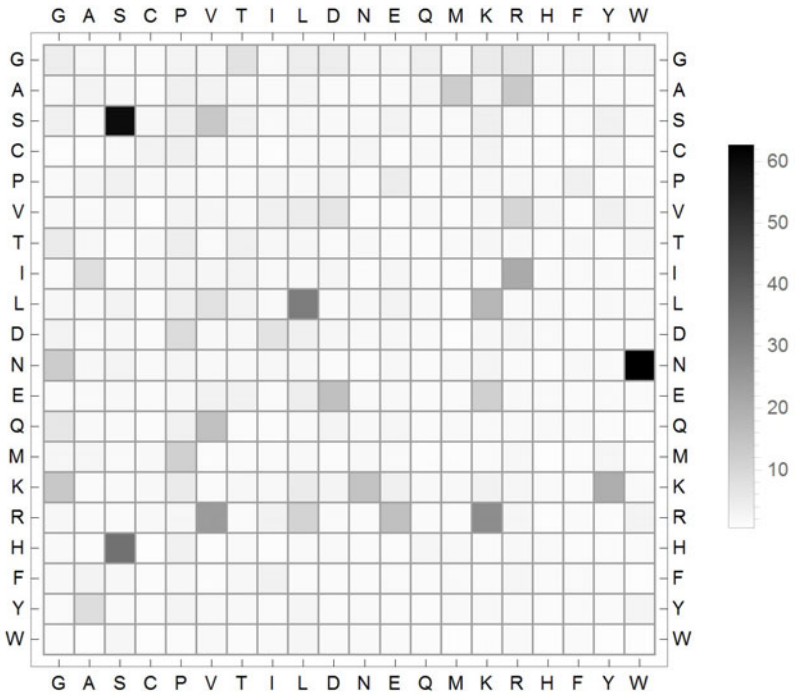

Fig. 5. Maximal volumes of the convex hulls of the peptide groups for all consecutive pairs of amino acid residues (the first residue is in row, the second one is in column) for all the NMR-files.

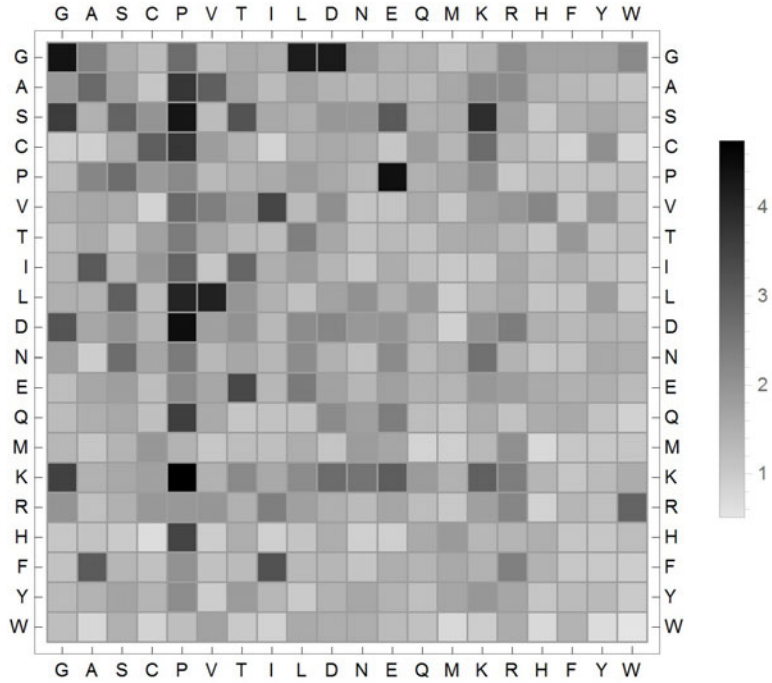

Fig. 6. Maximal volumes of the convex hulls of peptide groups for all consecutive pairs of amino acid residues (the first residue is in row, the second one is in column) for 2 836 selected files NMR-files.

One of the possible solutions is to use instead of the atom C_δ a virtual point on the ray $N - C_\delta$ at the distance from N equal to the mean value of the $N - H$ bond length. Another approach uses five atoms from a peptide group, namely one ignores the H atom in the second residue except proline and ignores C_δ for proline. In this section, we discuss the second approach to the plane law. Then the maximal volume of the convex hull of those five atoms for our selected files is $0.893A^3$. For volumes exceeding $0.5A^3$ one can clearly see significant deviation of five atoms from plane, while for volumes less than $0.4A^3$ the plane law is valid.

Let us see how pairs of consecutive amino acid residues with large deviation from the plane law are located with respect to the typical elements of the secondary structure such as helices and sheets. Note that the part of protein chain forming helices and sheets are indicated in PDB files (lines beginning with HELIX and SHEET, respectively).

To have an idea in which way the deviations from the plane law are distributed along a protein chain, we plot the following graphs. The number in the chain of a peptide bonds is on the horizontal axis (starting with the second one — recall that we do not consider the terminal amino acid residues). Along the vertical axis we plot the volume of the convex hull of the 5-tuples of atoms corresponding to the given peptide bond. The value of the volume is shown by a point, which is connected by a vertical line with the horizontal axis. The dotted line is for a peptide bond in which at least one of its atoms does not belong to any helix or sheet. The solid line is for two atoms of a peptide bond belonging to a helix, and the dashed line is for two atoms of peptide bond belonging to a sheet. Figure 7 is an example of such a graph plotted for the polypeptide containing the peptide bond with a maximum deviation from zero volume. This polypeptide has eight helices and nine sheets. Note that the largest deviations from the plane law are out of the helices and the sheets. Let us verify if this is a common rule. To do this, consider all the files that contain some deviation from the plane law greater than 0.5, and study in which files these deviations belong to a sheet or helix.

It appears that the number of files with a maximal deviation from the plane law exceeding the threshold 0.5 is equal to 45. Among these files only four have maximal deviation in some helix, and for

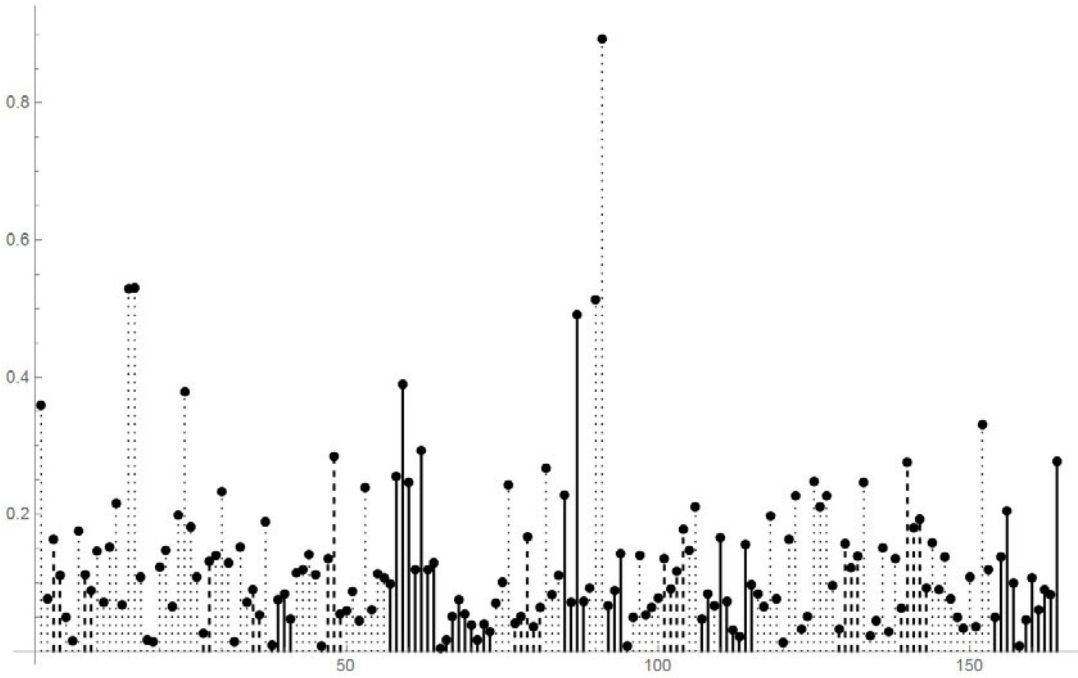

Fig. 7. Volumes for consecutive pairs of amino acid residues of the polypeptide 2JSY with helices and sheets marked.

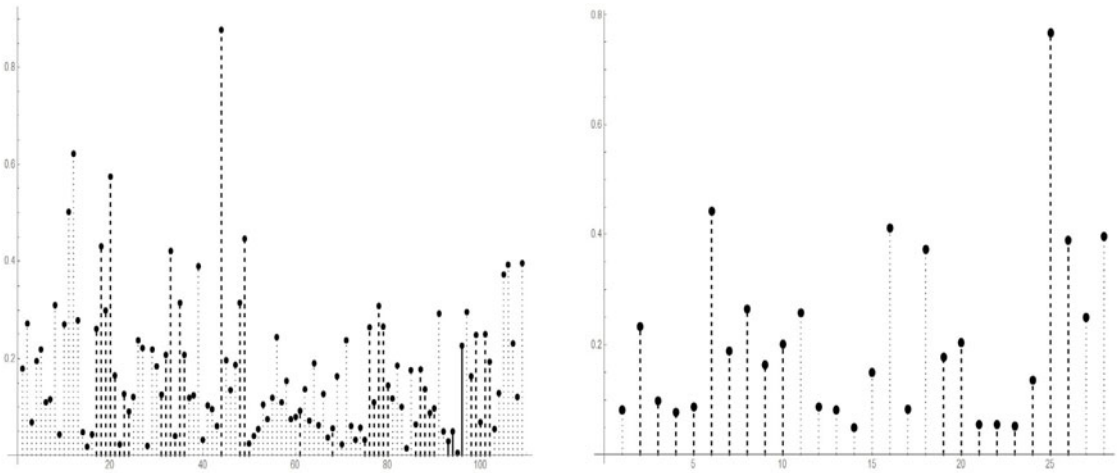

Fig. 8. Volumes for consecutive pairs of amino acid residues of 2H0P and 2NNT peptides: maximal deviations from the plane law belong to sheets.

eight of these files the maximal deviation belongs to a sheet. It is interesting that these two sets of files do not intersect.

The maximal deviation belonging to a sheet can be significantly big, for example, for protein 2H0P it is equal to 0.88, and for 2NNT it is equal to 0.77 (see Fig. 8). In five files from the eight files corresponding to the sheets, the maximal deviation is bigger than 0.62.

On the other hand, for helices the situation is different, namely the maximal deviation does not exceed 0.57. In Fig. 9, the example with maximal deviation (= 0.567) is shown. Note that in this example the peptide group with maximal deviation from the plane law belongs to the interior of the helix. In other three files, peptide groups with maximal deviations belong to the boundaries of helices (see Fig. 10).

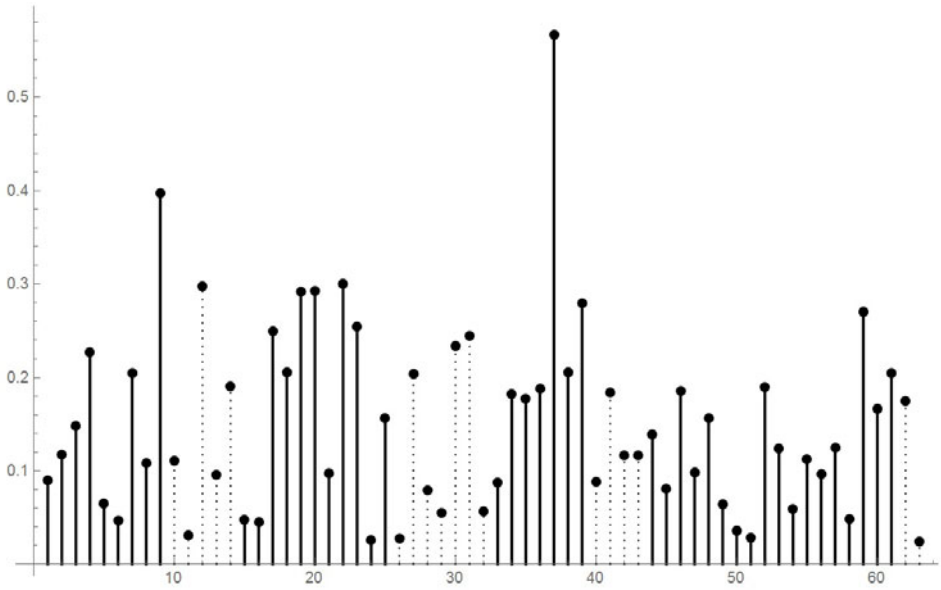

Fig. 9. Volumes for consecutive pairs of amino acid residues for the protein 1C5A: the maximal deviation from the plane law is inside a helix.

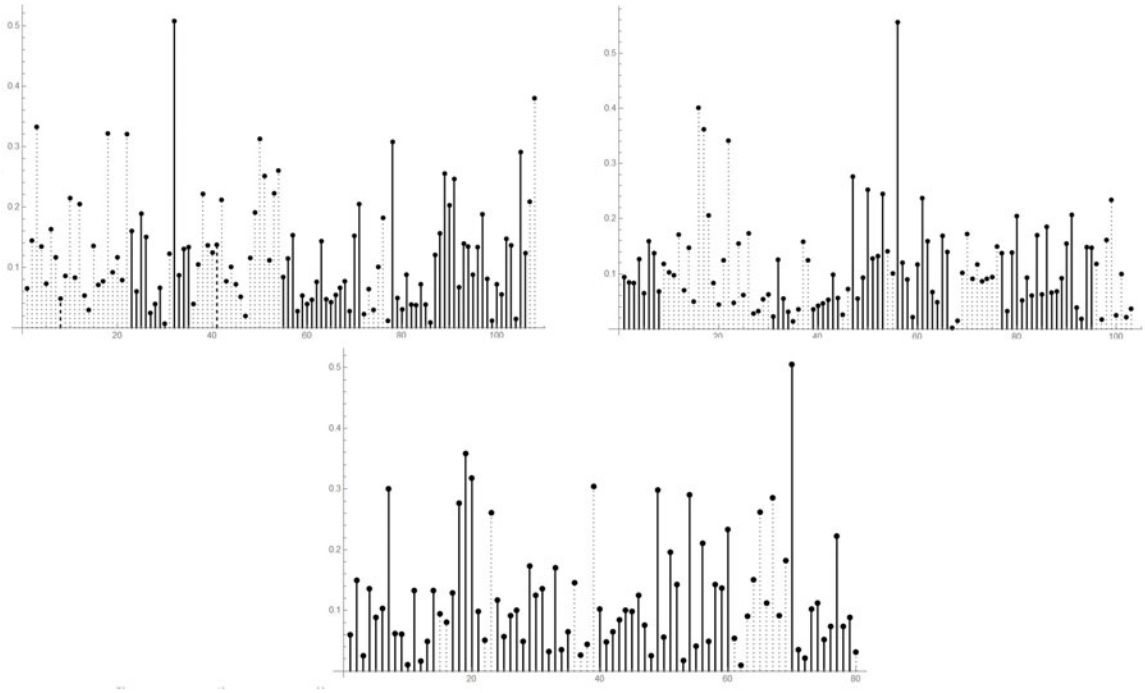

Fig. 10. Volumes for consecutive pairs of amino acid residues for the proteins 1XYJ, 2L4D, and 2LQL: the maximal deviation from the plane law is on the helix boundary.

Hence it seems natural to make a conjecture that helices and sheets as the most rigid parts of protein chains contain less deviations from the plane law than other parts. Also helices contain less deviations than sheets.

6. Conclusion Remarks and Acknowledgments

Our analysis allows us to make a conjecture that substantial deviations from the plane law are related to features of the protein secondary structure, namely the parts of the protein chain that do not form helices or sheets are more prone to such deviations. Also helices are more rigid than sheets. However, our approach works much better for the collection of files with small deviations of bond lengths and bond angles from their mean values than working on the full collection of files from the PDB (though those files also should pass certain tests before deposing the structure to the PDB). Our analysis shows that in many cases the PDB files may need to be corrected. Perhaps a more detailed analysis of optimizing procedures of the functionals that are used for finding protein structures from experimental data is needed.

The authors are grateful to M. Kirpichnikov and A. Fomenko for their support and attention to the work, and A. Arseniev, V. Golo, and K. Wüthrich for fruitful discussions. The work was financially supported by RScF, agreement No. 14-50-00029.

REFERENCES

1. A. Amunts, A. Brown, J. Toots, S. H. W. Scheres, and V. Ramakrishnan, "The structure of the human mitochondrial ribosome," *Science*, **348**, 95–98 (2015).
2. C. I. Branden and T. Jones, "Between objectivity and subjectivity," *Nature*, **343**, 687–689 (1990).
3. C. Hammond, *The Basics of Crystallography and Diffraction*, Oxford Univ. Press, Oxford (2009).
4. R. W. W. Hooft, G. Vriend, C. Sander, E. E. Abola, "Errors in protein structures," *Nature*, **381**, 272–272 (1996).

5. A. O. Ivanov, A. S. Mishchenko, and A. A. Tuzhilin, "Geometry of amino acids and polypeptides," *Nanostructures. Math. Phys. Model.*, **10**, No. 1, 39–47 (2014).
6. A. O. Ivanov, A. S. Mishchenko, and A. A. Tuzhilin, "Geometry of polygonal lines and polypeptides," *Nanostructures. Math. Phys. Modeling*, **10**, No. 1, 49–76 (2014).
7. A. O. Ivanov, A. S. Mishchenko, and A. A. Tuzhilin, "Critical analysis of amino acids and polypeptides geometry," in: V. A. Sadovnichiy and M. Z. Zgurovsky, eds., *Continuous and Distributed Systems: Theory and Applications*, Vol. 2, Springer, Berlin (2015), pp. 29–74.
8. *Protein Data Bank. An Information Portal*, `http://www.rcsb.org/pdb/home/home.do`.
9. *Protein Data Bank. An Information Portal. 2PDE*, `http://www.rcsb.org/pdb/explore.do?structureId=2pde`.
10. G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure*, Springer, Berlin (1979).
11. K. V. Shaitan, "The relaxation model of ideal folding in a homogeneous viscous medium," *Biophysics*, **60**, No. 5, 692–700 (2015).
12. K. V. Shaitan and I. A. Orshanskiy, "The molecular dynamics of the self-assembly and a rheological model of the superhelical structure of a spiderweb protofibril," *Biophysics*, **60**, No. 4, 538–541 (2015).
13. E. A. Vilkul and A. A. Tuzhilin, "Geometry of amino acids and polypeptides: the case of X-ray analysis," *Nanostructures. Math. Phys. Modeling*, **11**, No. 2, 5–27 (2014).
14. K. Wüthrich, "Protein recognition by NMR," *Nature Struct. Biol.*, **7**, 188–189 (2000).

E. A. Vilkul
Department of Probability Theory, Faculty of Mechanics and Mathematics,
Lomonosov Moscow State University, Moscow, Russia
E-mail: `elena.tuzhilina@mail.ru`

A. O. Ivanov
Department of Differential Geometry and Applications, Faculty of Mechanics and Mathematics,
Lomonosov Moscow State University, Moscow, Russia;
Department of Mathematical Modeling, Bauman Moscow Technical State University, Moscow, Russia
E-mail: `aoiva@mech.math.msu.su`

A. S. Mishchenko
Department of High Geometry and Topology, Faculty of Mechanics and Mathematics,
Lomonosov Moscow State University, Moscow, Russia
E-mail: `asmish.prof@gmail.com`

Th. Yu. Popelenskii
Department of Differential Geometry and Applications, Faculty of Mechanics and Mathematics,
Lomonosov Moscow State University, Moscow, Russia
E-mail: `popelens@mech.math.msu.su`

A. A. Tuzhilin
Department of Differential Geometry and Applications, Faculty of Mechanics and Mathematics,
Lomonosov Moscow State University, Moscow, Russia
E-mail: `tuz@mech.math.msu.su`

K. V. Shaytan
Department of Bioengineering, Biological Faculty,
Lomonosov Moscow State University, Moscow, Russia
E-mail: `shaytan49@yandex.ru`