ISSN 0006-3509, Biophysics, 2018, Vol. 63, No. 2, pp. 149–153. © Pleiades Publishing, Inc., 2018.
Original Russian Text © I.Yu. Torshin, A.V. Batyanovskii, L.A. Uroshlev, N.G. Esipova, V.G. Tumanyan, 2018, published in Biofizika, 2018, Vol. 63, No. 2, pp. 225–231.

Noncanonical and Strongly Disallowed Conformations of the Backbone in Polypeptide Chains of Globular Proteins

I. Yu. Torshin[a], A. V. Batyanovskii[b], L. A. Uroshlev[c], N. G. Esipova[c], and V. G. Tumanyan[c], *

[a]*Department of Chemistry, Moscow State University, Moscow, 119991 Russia*
[b]*Institute of Biophysics and Cell Engineering, National Academy of Sciences of Belarus, Minsk, 220072 Belarus*
[c]*Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991 Russia*
**e-mail: tuman@eimb.ru*

Received December 2, 2017

Abstract—An analog of the Ramachandran map was drawn, a new representation proposed, and thorough analysis performed using modern recognition and classification methods. Very large maps with a density of more than 50 million dots were created based on the data sets derived from the latest releases of globular protein-structure data banks. A, B, B', C, and D regions that correspond to strongly disallowed conformations were defined and found to occupy 25% of the plot area. A region of noncanonical conformations was determined by subtracting strongly disallowed and permitted conformation regions from the total plot area. Arguments are provided to support the new classification of backbone conformations of the protein polypeptide chain.

Keywords: globular proteins, Ramachandran map, conformational analysis, recognition and classification, statistical and cluster analyses, strongly disallowed, permitted, noncanonical conformations

DOI: 10.1134/S0006350918020240

INTRODUCTION

Ramachandran et al. [1, 2] proposed a two-dimensional map as a method to describe and analyze the backbone conformations of a polypeptide chain because the only two parameters per residue are the main conformational parameters that determine the backbone structure (five parameters per structural unit are necessary in the case of a polynucleotide chain). The two parameters are the dihedral angle φ between the planes through the atoms C', N, and C^{α} and N, C^{α}, and C' and the dihedral angle ψ between the planes through the atoms N, C^{α}, and C' and C^{α}, C', and N. The angle ψ between the planes through the atoms C^{α}, C', and N and C', N, and C^{α} may be considered constant in the first approximation owing to a basic property of the polypeptide bond; i.e., sp^2 hybridization of the nitrogen atom of the amino group and the carbon atom of the carbonyl group renders the peptide bond virtually planar. When the space of natural parameters is used in place of the Cartesian coordinates of atoms to describe the conformation, the saving in the number of independent variables is especially great in the case of polypeptide chains because only two parameters define the so-called dipeptide unit, which includes at least 13 atoms. Because there is a one-to-one correspondence between protein-backbone conformations and pairs of the dihedral angles φ and ψ, a conformation map obtained as a function of the two independent variables has become the gold standard in the conformational analysis of proteins.

Conformational analysis is performed to achieve two goals, to identify the permitted conformations for a molecule or its segment and, on the other hand, to identify its disallowed conformations. More favorable energies are thought to correspond to the former; and less favorable energies, to the latter. Apart from these, intermediate conformations are identified in actual analyses and belong to the so-called gray zone, whose interpretation of which may vary.

Early conformation maps with permitted regions outlined and experimental (φ, ψ) pairs plotted have already made it possible to characterize both permitted and disallowed regions [1, 2]. An analysis of the maps [1, 2] leads to the fundamental conclusion that the permitted regions are approximately four times greater in area than the disallowed regions on a map of a dipeptide unit. The proportion and outlines of the permitted and disallowed regions are naturally refined with progress in computational and experimental research.

For the permitted regions, analyses of experimental protein structures in accordance with computational data have revealed two distinct permitted clusters, which correspond to α-helices and β-structures, and a minor cluster, which corresponds to left-handed α-helices (Fig. 1). Further studies have shown that the

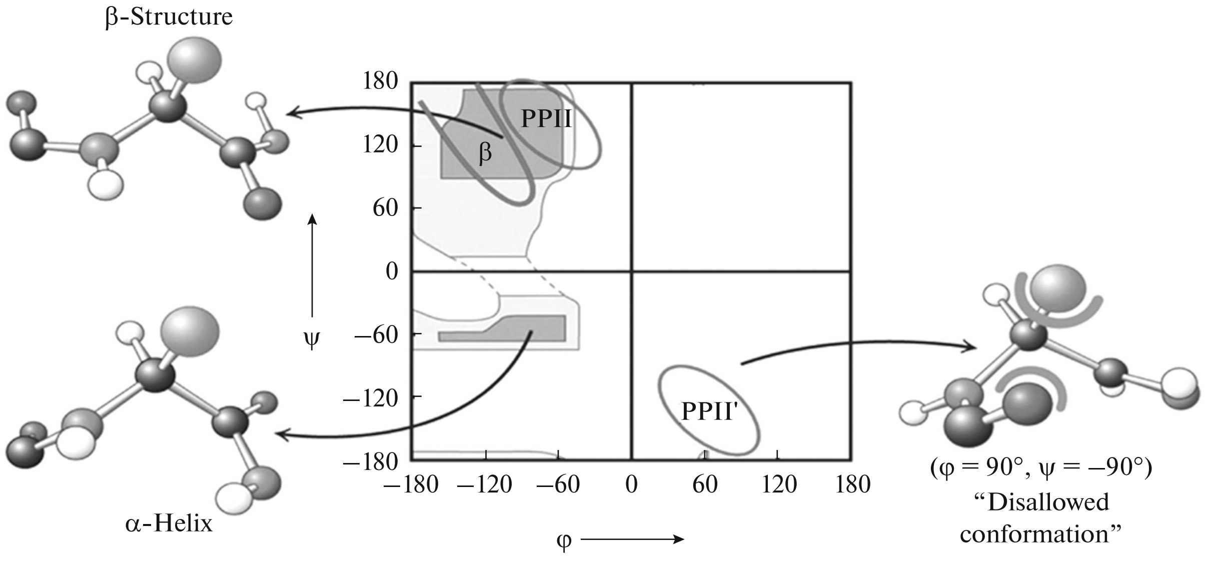

Fig. 1. The conformations of the backbone in polypeptide chains. A schematic Ramachandran plot is shown together with regions and structures that correspond to α-helices, β-structures, and PPII (left-handed polyproline II conformations [3, 4]). An example of disallowed, or forbidden, conformations is given (see text for comments).

left upper quadrant harbors two, rather than one, clusters of points; i.e., a region of left-handed polyproline II (PPII) conformations has been identified in addition to the region of β-structures [3, 4].

A disallowed, or forbidden, conformation is shown as an example in Fig. 1 together with permitted regions. The conformation corresponds to a type II' β-turn or conformation PPII', which is symmetrical to PPII, according to a nomenclature [5, 6].

The problem of disallowed regions of the Ramachandran plot requires detailed consideration. Early X-ray studies and experimental solutions of the structure for a relatively small number (several tens or hundreds) of proteins have already shown that, in addition to permitted regions, the Ramachandran plot includes regions that remain virtually empty in accordance with steric limitations characteristic of the polypeptide chain in the corresponding conformation. Single experimental points that occasionally occurred in the sterically disallowed regions were naturally interpreted as experimental errors or a phenomenon that requires special investigation. As protein structures accumulated in data banks and their quality (resolution) improved, experimental errors were rejected to leave conformations that were reliable although they fall in the disallowed regions.

Once established as a fact, such conformations were termed conformationally disallowed because they correspond to regions of conformations that are less favorable sterically compared, for instance, with regions that correspond to α-helices or β-structures (Fig. 1). Several slightly different variants were proposed for a definition of disallowed regions [7, 8]. In some studies, the term "disallowed conformation" is avoided, and sparsely populated zones are isolated to include the conformations that were previously classified as disallowed [9]. Thus, a purely phenomenological term that reflects the occurrence of respective conformations tends now to be used in place of the term that implies energy evaluation of conformations. It should be noted that terminology based on statistical evaluations appears to be more correct. In fact, the energy is difficult to estimate for a local structure because a role is played not only by its context, but also by factors that are scarcely controllable, such as those of a quantum-chemistry nature.

Looking ahead, our objective was to obtain the arguments that support division of the Ramachandran plot into regions of favorable energies, strongly disallowed regions, and all other regions, which we believe are reasonably called noncanonical. The study was based on an analysis of the latest PDB releases and new methods of point clustering, in our case, in the conformation space.

One of the reasons that we revisited the problem of the Ramachandran plot is that the structures that are newly deposited in the PDB are growing in number at an increasing rate and are improving in quality. The last comprehensive revision of the Ramachandran plot was reported in [5, 6]. A study [6] published in 2010 analyzed a set of high-resolution structures (≤1.2 Å) with data on 72000 amino-acid residues. In this work, we used the PDB release of 2016 and analyzed more than 50 million amino-acid residues. Moreover, new methods have been developed to analyze large data sets. Thus, we constructed, analyzed, and interpreted a Ramachandran plot on the basis of the experimental data that have accumulated to date and new data-processing techniques in this work.

What are the questions that are possible to answer via conformational analysis using the Ramachandran plot? As in the general case, conformational analysis makes it possible to identify the preferable conformations and to explain the causes of their stabilization. On the other hand, conformational analysis makes it possible to define the range of conformations that are disallowed to a particular extent and to study the nature of the relevant restrictions.

To analyze large data sets with respect to the angles (φ, ψ), it is of interest to employ new clustering methods that, first, are suitable for processing large data sets (in our study, a pair of the angles φ and ψ corresponds to each point; tens of millions of points occur) within a reasonable period of time and, second, have higher sensitivity for detecting clusters of points. An approach based on the concept of metrics was employed in cluster analysis in this work (in mathematics, a metric is a positive definite symmetrical function that measures distances in pairs of points and satisfies the triangle inequality). Measuring pairwise distances between the points allows metric clustering, that is, the identification of clusters of closely packed points with high point densities [10]. Likewise, the method identifies the regions where points occur at a low density; these regions correspond to disallowed regions.

MATERIALS AND METHODS

Data set. The PDB release of 2016 was analyzed. Files with identical sequences and resolutions lower than 2.0 Å were eliminated. The resulting set included the structures of 121450 protein chains, which were retrieved from 62096 PDB files. The angles (φ, ψ) were calculated for 52563104 amino-acid residues with coordinates known for each non-hydrogen atom of the backbone.

Clustering methods to identify high- and low-density regions. Algorithms to identify metric clusters (groups of closely packed points with high point densities) in a set of points with a given metric (the so-called metric configurations) and their strict mathematical grounding have been described previously [10]. Our experiments with model clusters that differ in their extent of smearing showed that our clustering procedure identifies clusters even in the case of minor fluctuations in point density. The clusters identified by the procedure cannot be identified using standard algorithms, such as DBSCAN, OPTICS, DeLi-Clu, and EM-clustering.

RESULTS AND DISCUSSION

Plotting the angles (φ, ψ) of 52563104 amino-acid residues on a Ramachandran map revealed four distinct low-density regions, which presumably correspond to the conformations that are avoided. The regions were designated A, B and B', C and C', and D (Fig. 2). We propose that the regions be called strongly disallowed because only a small number of amino-acid residues had conformations that correspond to the regions even in our set of more than 50 million points. The φ and ψ ranges that correspond to the regions are evident from Fig. 2. It is important that the result is maximally general because we did not perform any selection with respect to amino-acid composition, local-structure type, polypeptide-chain stereochemistry, etc.

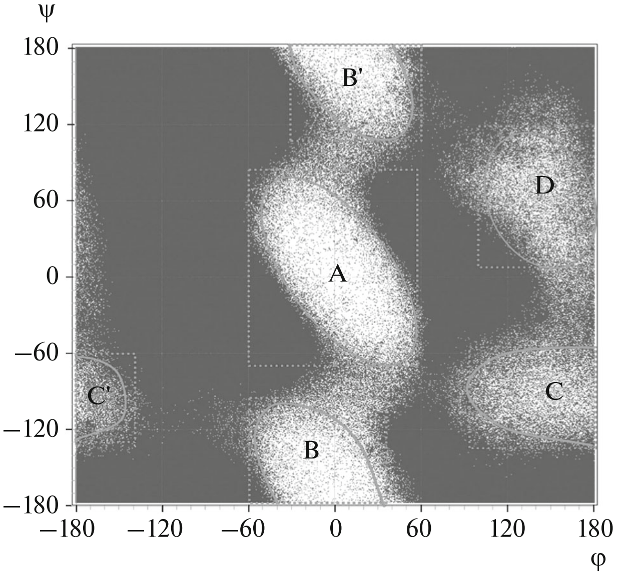

Fig. 2. The strongly disallowed regions on the Ramachandran plot obtained for 52 563 104 amino-acid residues are regions with the lowest point density. Four low-density regions are clearly seen. Because these regions lack distinct borders, ovals (solid lines) show the area that includes more than 90% of the points for each low-density region. Straight (dashed) lines approximate the low-density regions with rectangles in terms of the angles φ and ψ.

The positions of the strongly disallowed regions on the Ramachandran plot (Fig. 2) were compared with the published data. Commonly accepted data on atom–atom contacts of backbone atoms that occur on the Ramachandran plot were taken from [11], and additional data were obtained from a more recent work [12].

Atoms are conventionally numbered a certain way in the so-called Ramachandran dipeptide unit. Amino-acid residue atoms at the center of the unit lack indices in designations; atoms of the left adjacent carbonyl group of the previous residue have the index $i - 1$; and those of the amino group of the next residue have the index $i + 1$.

The contact $O_{i-1}...H_{i+1}$ [11] corresponds to strongly disallowed region A. The contact $O_{i-1}...N_{i+1}$ has been proposed in place of $O_{i-1}...H_{i+1}$ in [12]

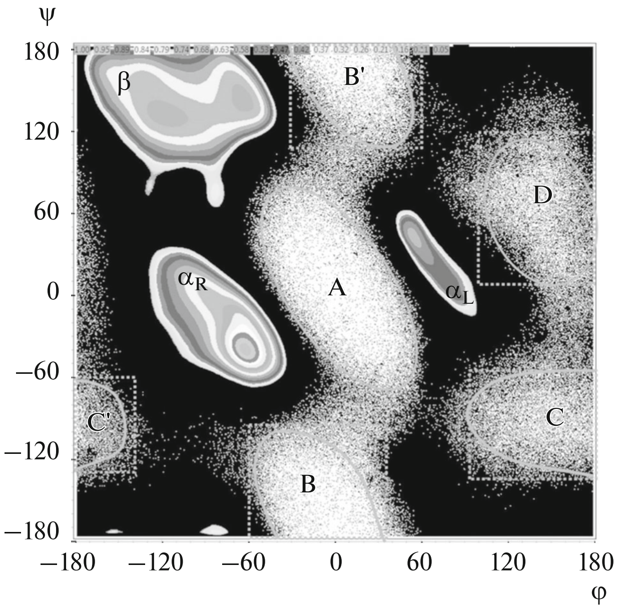

Fig. 3. Point clusters on the Ramachandran map. A color-coded scale (at the top) is used to show generalized density values. Cluster β corresponds to β-structure and PPII regions; cluster α_R, right-handed α-helices; and cluster α_L, left-handed α-helices. Density peaks in the region φ = −180°...−60°, ψ = −180°...170° can be considered as part of cluster β. Low-density regions are also shown on the map in white, which corresponds to a zero point density in the respective area of the map. Strictly speaking, exact outlines of high-density regions depend on the amino-acid identity and modulation of the angle ω, which defines the deviation from the plane of the peptide group, as has been demonstrated in [18].

(regions of the contacts are bounded with the ellipse). A contribution to region A is made by the contacts C_{i-1}...C (a vertical band on the map) and N...H_{i+1} (a horizontal band on the map).

For regions B and B', the characteristic contacts are O_{i-1}...O (bounded by the ellipse) and C_{i-1}...C (the vertical band on the map), which is the same as in the case of region A.

The interpretation of region C in terms of contacts can be associated with the horizontal band that corresponds to the contacts C^β...H_{i+1} (or C^β...N_{i+1} in [12]) and the vertical band of the O_{i-1}...C^β contacts.

Finally, the existence of region D can be explained by simultaneous superimposition of the contacts O_{i-1}...C^β and N...H_{i+1} (and C^β...O according to [12]) and overlap of the Van der Waals radii of H...H_{i+1}.

As can be seen, the interpretation of regions D and C in terms of contacts is not as convincing as that of regions A and B, as the shapes of the region are considered, while the restrictions are also somewhat weaker in the case of regions D and C.

It should be noted that our results are not feasible based on strained Van der Walls contacts because these contacts occur in both strongly disallowed and noncanonical regions. Such a simplified concept does not allow the diagonal shapes of the disallowed regions or substantial gaps between them. In the case of regions A and B, the gaps agree well with the diagonal bands that occur between them and correspond to the stabilizing $n \rightarrow \pi^*$ interaction, which is a purely quantum-mechanical effect (see Fig. 3 in [13]).

Once the low-density regions are reliably established, apparent high-density regions can be considered. Our analysis of point clusters on the Ramachandran map (Fig. 3) employed a highly sensitive metric clustering method and revealed three high-density regions. The first one corresponded to β-structures and left-handed helices (Fig. 1, region PPII), while the second and third ones corresponded to right- and left-handed α-helices, respectively. No other cluster was detected; all other regions of the map were uniformly filled with points to low point densities (η_j < 0.03) and lacked isolated density peaks.

Thus, if the strongly disallowed regions of the Ramachandran map are identified as regions with the lowest, almost zero point density (the above low-density regions), then regions A, (B, B'), (C, C'), and D are such regions in our case (Fig. 2). The total area of these regions is rather great, accounting for approximately 25% of the total map area. The definition of strongly disallowed regions is based on modern sets of protein structures that were solved to a resolution of not worse than 2 Å. If we combine these regions with the permitted conformation regions (apparent high-density regions, Fig. 3) and subtract their sum from the total Ramachandran map (Fig. 4), the remaining regions will correspond to the so-called disallowed conformations [7, 8, 15–17], or conformations of sparsely populated zones according to another classification [9]. We prefer the term “noncanonical conformations” here.

Figure 4 shows the disallowed conformation regions as they have been defined in [7, 8]. It is seen that the regions agree well with the conformation-map regions that are neither strongly disallowed nor permitted according to our terminology. This conclusion remains true even though the results of [7, 8] are not fully coherent.

Thus, the results of the localization of conventionally disallowed regions, which have been proposed in earlier studies, and strongly disallowed and permitted regions indicate that regions of the Ramachandran map should be classed into three categories: permitted regions (which correspond to the density peaks in Fig. 3), strongly disallowed regions (the low-density regions in Fig. 2), and noncanonical regions. Regions of the last category are similar to a certain extent to the conventionally disallowed regions [7, 8], which are not strongly disallowed, but are not permitted as well.

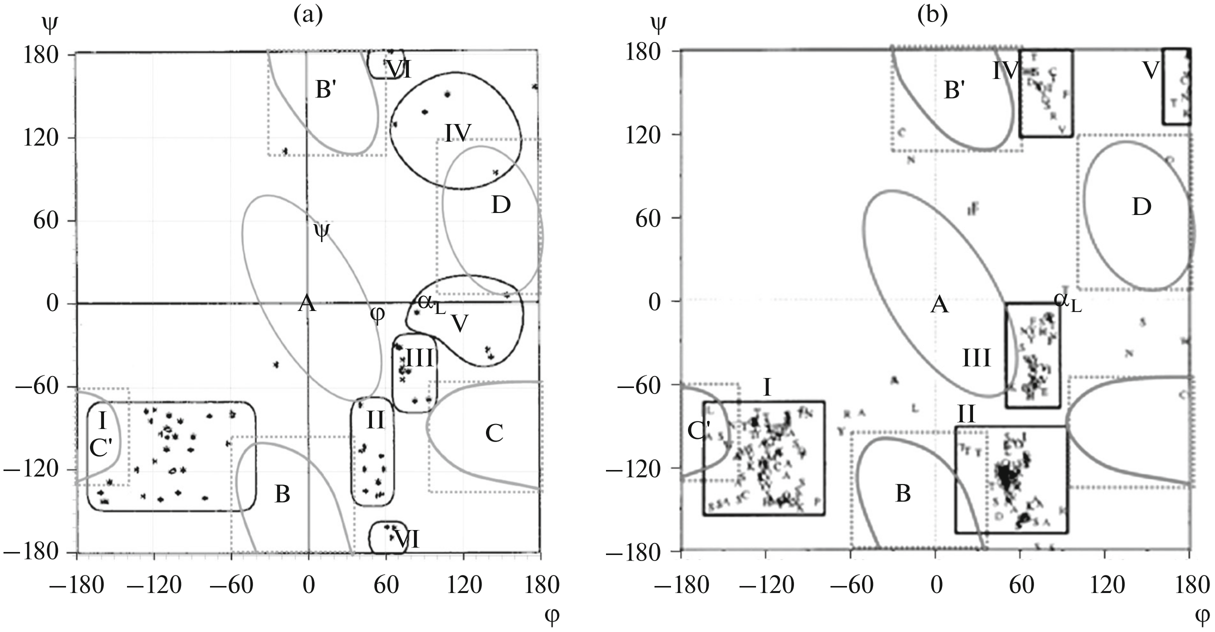

Fig. 4. Comparison of the positions of low-density regions with those of the conventionally disallowed regions identified in (a) [7] and (b) [8].

These regions join with the sparsely populated zones [9].

All regions of the Ramachandran map that are not permitted or strongly disallowed should be classified as conventionally disallowed. The problem is that the difference in point density between high- and low-density regions should exceed 5–7% for a point cluster to be identified by the procedure used in this work. With this accuracy, it is not feasible to reliably distinguish the point clusters in regions I, II, and II'.

The identities of amino-acid residues were disregarded in our analysis of backbone conformations. Backbone-conformation patterns of individual amino acids have been considered elsewhere [22].

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project nos. 17-04-02105 and 16-54-00219-Bel) and the Belarussian Republican Foundation for Basic Research (project no. B16R-178).

REFERENCES

1. G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, J. Mol. Biol. **7**, 95 (1963).
2. C. Ramakrishnan and G. N. Ramachandran, Biophys. J. **55**, 909 (1965).
3. A. A. Adzhubei, M. J. Sternberg, and A. A. Makarov, J. Mol. Biol. **425**, 2100 (2013).
4. N. G. Esipova and V. G. Tumanyan, Curr. Opin. Struct. Biol. **42**, 41 (2017).
5. S. A. Hollingsworth, D. S. Berkholz, and P. A. Karplus, Protein Sci. **18**, 1321 (2009).
6. S. A. Hollingsworth and P. A. Karplus, Biomol. Concepts **1**, 271 (2010).
7. K. Gunasekaran, C. Ramakrishnan, and P. Balaram, J. Mol. Biol. **264**, 191 (1996).
8. D. Pal and P. Chakrabati, Biopolymers **63**, 195 (2002).
9. N. V. Kalmankar, C. Ramakrishnan, and P. Balaram, Proteins **82**, 1101 (2014).
10. I. Yu. Torshin and K. V. Rudakov, Pattern Recogn. Image Anal., No. 4. 145 (2016).
11. N. Mandel, G. Mandel, B. L. Trus, et al., J. Biol. Chem. **252**, 4619 (1977).
12. B. K. Ho, A. Thomas, and R. Brasseur, Protein Sci. **12**, 2508 (2003).
13. M. P. Hinderaker and R. T. Raines, Protein Sci. **12**, 1188 (2003).
14. I. Yu. Torshin, A. V. Batyanovskii, L. A. Uroshlev, et al., Biophysics (Moscow) **62** (3), 342 (2017).
15. M. C. Vega, J. C. Martinez, and L. Serrano, Protein Sci. **9**, 2322 (2000).
16. L. A. Uroshlev, I. Yu. Torshin, A. V. Batyanovskii, et al., Biophysics (Moscow) **60** (1), 1 (2015).
17. I. Yu. Torshin, N. G. Esipova, and V. G. Tumanyan, J. Biomol. Struct. Dynam. **32** (2), 198 (2014).

Translated by T. Tkacheva