The Ramachandran Map of More Than 6,500 Perfect Polypeptide Chains

Zoltán Szabadka,

Rafael Ördög,

Vince Grolmusz

manuscript received March 19, 2007

Z. Szabadka, R. Ördög and V. Grolmusz are with Eötvös University, H-1117 Budapest, Hungary

Abstract

The Protein Data Bank (PDB) [3] is the most important depository of protein structural information, containing more than 42,000 deposited entries today. Because of its inhomogeneous structure, its fully automated processing is almost impossible. In a previous work, we cleaned and re-structured the entries in the Protein Data Bank, and from the result we have built the RS-PDB database [11]. Using the RS-PDB database, here we draw a Ramachandran-plot from 6,593 "perfect" polypeptide chains found in the PDB, containing 1,192,689 residues, this is more than tenfold increase in the size of data analyzed before this work. The density of the data points makes possible to draw a logarithmic heat map enhanced Ramachandran map, showing the fine inner structure of the right-handed α-helix region.

Keywords: Protein Data Bank, biochemical databases, database cleaning, Ramachandran map, torsion angles, multi-protein Ramachandran plot

I. Introduction

The Protein Data Bank [3] developed as the publicly available depository of the three-dimensional structural data of proteins and nucleic acids. As long as most researchers were interested in one protein, it was perfectly adequate to use human assisted pre-processing of that pdb-file for review or structural studies; frequently even the accompanying journal publication had to be reviewed for proper interpretation of the data deposited. Today, however, the PDB contains more than 42,000 entries, and lots of studies would use this treasury, involving even hundreds or thousands of proteins.

Unfortunately, the inhomogeneous structure of the PDB makes the completely automated processing difficult. The PDBsum pictorial database [7], [1] contains lots of improvements over the original PDB, but it is mainly a graphical interface to the original PDB. In [12] we presented an algorithm for reliably finding protein-ligand complexes in the Protein Data Bank, and also for repairing certain inconsistencies in the database.

Seeking methods for arranging protein structural data by strict mathematical rules, we further developed the ideas from [12]: we have built the Rich-Structure PDB (abbreviated RS-PDB) database from the PDB data [11]. That database contains the result of a mathematical procedure in which every peptide- or nucleotide-chain and ligand molecule were re-built, verified and classified by completely automatic processing using basically graph-theoretic algorithms [11].

In [6], using the RS-PDB database, we identified protein-ligand binding sites and analyzed the residue-composition on the binding sites from the whole Protein Data Bank.

In the present work we intend to analyze the torsion angles of the peptide backbone

in those peptide-chains in the PDB which satisfy some strict quality requirements; however, we do not restrict our study to some small subset of data: from the strict structure of the RS-PDB database one is able to choose and process thousands of peptide-chains for such a study.

Our goal was to draw a Ramachandran map [9] of as large high-quality data set as possible from the PDB.

The Ramachandran map is an unparalleled tool in protein structure validation and prediction, and the better understanding of its properties and the shapes of its regions is an active research field today (e.g., [10], [5], [8]).

As reported in the literature, polypeptide chains are usually chosen by the resolution of the underlying PDB entry as the quality criterion; the numerous redundancies in the PDB [12] are usually addressed by some constraints on the residue sequence.

For the drawing and the analysis of higher-order Ramachandran plots in [10], PDB entries with better than 1 Å resolution and less than 25% sequence similarity were chosen: the data set contained 51 structures and 10,976 $\phi - \psi$ pairs.

In the deep analysis of [5], a dataset of [8] was used, consisting of 500 non-homogeneous structures of resolution better than 1.8 Å; the set contained 97,368 residues.

In our present work, we use more intricate quality criteria and found 6,593 pairwise different high quality polypeptide chains, containing 1,192,689 residues. The Ramachandran map of our data-set is given on Figure 1.

II. METHODOLOGY

We used the June 1, 2006 version of the PDB for our work.

We intended to process as many polypeptide chains as it was possible without sacrificing data validity. Several authors filtered the PDB data by

First, we have chosen those polypeptide chains, where all the atomic coordinates of the heavy (i.e., non-hydrogen) atoms were given in the PDB mmCIF file. We verified this by comparing the entries of the residue-sequence information with the atomic coordinate-data. This step was done since missing atoms either in the polypeptide chains or in the side-chains would imply one or more incorrect $\phi-\psi$ angle readings.

Next, the residue-composition of the polypeptide chains were reviewed. We have chosen only those chains where no modified residues were present: modified residues, due to their different geometry, would imply different $\phi-\psi$ angle distributions. For the list of modified residues found in the PDB, see Table 2 in [12].

Next, we verified the correctness of the lengths of the covalent bonds in the polypeptide chains. In the process of building the RS-PDB database [11], [6], we generated a table, containing bond errors, as follows:

- By building a kd-tree [2], all the atom-pairs of distance less than 6 Åwere identified.
- If two atoms from different residues or from different polypeptide-chains were placed within covalent distance, we enter this fact to the bond-error table.
- If two atoms from the same residue were placed within covalent distance, but between them no covalent bond may exist, then we also enter this fact to the bond-error table.

We use only those polypeptide chains for constructing the Ramachandran-plot, what do not contain bond errors, listed in the database table.

As the last step, we filtered out the homogeneous sequences by using an easily applicable hash [4] function (an MD5 hashing), and the remaining set consisted of pairwise non-homogeneous polypeptide chains.

After this filtering procedure, we were left with 6,593 polypeptide chains containing 1,192,689 residues.

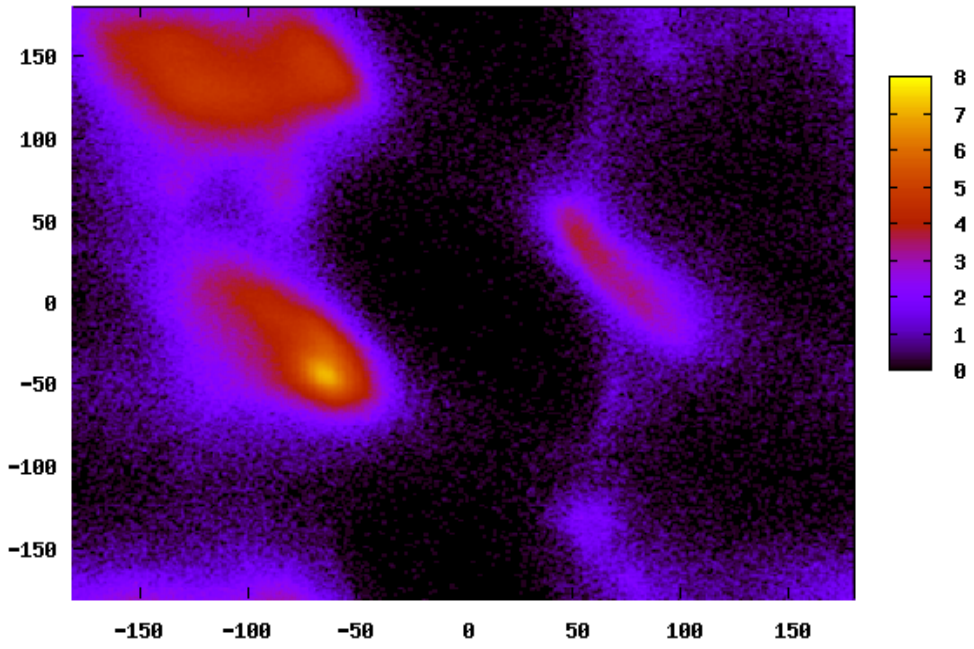

Fig. 1. The logarithmic heat-map enhanced Ramachandran plot of 1,192,689 residues. The colors correspond to the natural logarithm of the density of the data-points in the map. Note the density distribution on the right-handed α-helix region.

III. Results and discussion

Because of the large size of our data-set, we were able to draw a logarithmic heat-map enhanced Ramachandran plot [9] from 6,593 polypeptide chains containing 1,192,689 residues on Figure 1. The colors correspond to the natural logarithm of the density of the data-points in the map. Note the density distribution in the right-handed α-helix region, close to the point (-60,-50): using the color-coding, one get clearly quantitative distribution of the densities.

By the best of our knowledge, no such heat-map enhanced Ramchandran-plot appeared before in the literature, since the data sets were clearly too small.

REFERENCES

[1] Laskowski R. A., Chistyakov V. V., and Thornton J. M. Pdbsum more: new summaries and analyses of the known 3d structures of proteins and nucleic acids. *Nucleic Acids Res.*, 2005.

[2] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[4] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to algorithms*. MIT Press, Cambridge, MA, second edition, 2001.

[5] Bosco K. Ho, Annick Thomas, and Robert Brasseur. Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the α-helix. *Protein Science*, 12:2508–2522, 2003.

[6] Gabor Ivan, Zoltan Szabadka, and Vince Grolmusz. Cysteine and tryptophane anomalies found when scanning all the binding sites in the Protein Data Bank. In *submitted*, 2007.

[7] R. A. Laskowski. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Research*, 29:221–222, 2001.

[8] S.C. Lovell, I.W. Davis, W.B. Arendall III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson, and D.C. Richardson. Structure validation by C_α geometry: φ, ψ and C_β deviation. *Proteins*, 50:437–450, 2003.

[9] G. N. Ramachandran and V. Sasisekharan. *J. Mol. Biol.*, 7:95–99, 1963.

[10] Gregory E. Sims, In-Geol Choi, and Sung-Hou Kim. Protein conformational space in higher order φ-Ψ maps. *Proc. Nat. Acad. Sci.*, 102:618–621, 2005. doi:10.1073/pnas.0408746102.

[11] Zoltan Szabadka and Vince Grolmusz. Building a structured PDB: The RS-PDB database. *Proceedings of the 28th IEEE EMBS Annual International Conference, New York, NY, Aug. 30-Sept 3, 2006*, pages 5755–5758, 2006.

[12] Zoltan Szabadka and Vince Grolmusz. High throughput processing of the structural information in the Protein Data Bank. *J. Mol. Graph. Model.*, 25(6):831–836, 2007.