Statistical Analysis of Protein Side-chain Conformations

Andrei Ignatov
HSE University, Moscow, Russia
Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia
E-mail: rayignatov@outlook.com

Abstract. In the paper, three algorithms for predicting protein side-chain conformations are suggested and discussed. All proposed approaches analyze the local neighborhood of the target residue to avoid 'steric clashes'. Strong and weak points of the algorithms are described, and ways of improving their outcomes are suggested. The approach based on predicting conformations for all residues in a protein chain segment appears to be the most promising.

1. Introduction

Experimental methods of protein structure determination are still challenging and require huge costs both in time and budget. That is why computational methods of structure modelling are considered an important task.

Most of modern structure prediction tools are focused on correct backbone determination as they use coarse-grained models [1, 2, 3, 4] or $C_\alpha$ or $C_\beta$ pairwise distances [5] for global optimization. These instruments help in understanding the general organization of structure but do not fit the real needs of computational biology research tasks. Instead, problems of protein-protein docking or free energy calculation depend on accurate all-atom structure. Thus, the mentioned tools are hardly applicable to these problems.

Several force fields like AMBER, CHARMM, or OPLS [6, 7, 8] can perform all-atom minimization of structure using fast and reliable gradient-based methods. However, the energy function of macromolecule is quite nonconvex, so any local optimization method strongly depends on the initial conformation guess [9]. Therefore, it may be concluded that the closer is the initial guess to the target extrema point, the faster the optimization process will converge. As a result, it is necessary to predict side-chain conformation with a confidence level that would be high enough to guarantee fast minimization. Several assumptions are required to perform this guess.

First of all, we assume that lengths of covalent bonds are approximately constant, so we can use a rough approximation of structure description in terms of dihedral angles. Then, we can consider only some stable dihedral states that are used at most while all other states are poorly presented. Finally, we assume that we can test different side-chain conformations keeping the backbone fixed.

In this research, we examine protein structures to figure out special inter- and intra-amino acid dependencies that influence the conformation of side-chains. We test several approaches...
to predict $\chi_i$, $i = 1, 5$ dihedral angle values of side-chains assuming that backbone structure is provided.

Similar work was performed in [10] where side-chain dihedral angles were predicted by coordinates of backbone atoms. In [11] authors describe another method based on graph theory where two rotamers with nonzero interaction energy are represented as ends of an edge. Interaction graphs were also used in [12] where protein structure was modelled as a Markov random field. Another article [13] describes a packing optimization algorithm applied to van der Waals interactions. However, the authors emphasize that these results either lack accuracy or work slowly, so we decided to apply relatively fast Machine Learning (ML) algorithms to this problem. All our methods are intended to make an initial prediction based only on the protein structure.

In the following sections, the examined approaches are discussed.

Most of the code for this research was written in Python programming language. As an information source, we decided to use the Protein Data Bank (PDB) base [14] and analyze its files with Biopython module [15].

2. Basic Rotamer Prediction

Analysis of dihedral angle values shows that side-chain conformation is likely to be influenced by the local context of the target amino acid residue. This implies that some characteristics of neighbor residues should be taken into account somehow.

We define neighborhood as the distance between $C_{\alpha}$ atoms smaller than some fixed threshold. It should be noticed here that residues may be close to each other spatially but far from each other along the chain. In this approach, we selected 15 Å as the cutoff value. Having that, it was discovered that 17 neighbor residues are likely to be present in this radius with 0.97 confidence.

Thus, in this approach, we consider features of neighbor residues and predict side-chain conformation for every amino acid type separately. Speaking of predicted values themselves, side-chain rotamers are considered. A rotamer is one of several (mostly three) stable positions of side-chains that are created by rotation around one chemical bond. Rotamers correspond to the local minimums of potential energy of the protein. They can be seen as peaks at histograms of dihedral angle values. For example, Figure 1 presents the histogram of $\chi_1$ angle values for glutamine acid ($\text{GLU}$). However, not all degrees of freedom appear to be rotameric: some angles have ‘messy’ distributions like the one presented in Figure 2. We resolve this issue by analyzing dependencies between non-rotameric $\chi_{i+1}$ distributions and $\chi_i$ rotamers. After analyzing $(\chi_i, \chi_{i+1})$ pairs, we find that for some stable values of $\chi_i$, values of a non-rotameric $\chi_{i+1}$ look similar to a rotameric distribution. This fact was used to figure out the non-rotameric peaks. An example of such approach is presented in Figure 3.
For retrieving data for this research, we first filtered the PDB base. Currently, Protein Data Bank consists of more than 140 thousands of files. Besides files representing single proteins, the data contains short chains (peptides) and protein complexes. Files also may have gaps in atoms or even in residue sequences. Thus, there is a clear need for filtering. For our purposes, we selected the files containing one chain; consisting of at least 100 residues; having only $\alpha$-amino acid residues; and having no missed residues and atoms.

After PDB filtering, a set of appropriate PDB id’s was formed. From these files, data regarding every side-chain dihedral angle $\chi_i$, $i = 1, 5$ was collected. Such datasets were collected separately for every amino acid except for alanine ($ALA$, $A$) and glycine ($GLY$, $G$) due to the small length of their side-chains. The collected sets include the following features of neighbor residues (for each sample):

(i) Values of backbone dihedral angles $\phi, \psi$;
(ii) Distance to the target residue (in terms of $C_\alpha$, measured in Å);
(iii) Physicochemical features [16, 17, 18];
(iv) Polarity groups;
(v) Several geometrical features.

After collection, the datasets were provided to ML algorithms. We conducted several experiments to study the results of different Machine Learning approaches. LightGBM algorithm (based on decision trees) appeared to show the best results that are presented in Table 1.

The results achieved with LightGBM are only slightly better than the outcomes of the Rotamer Library prediction method [19]. It may be explained by the fact that our approach considers only the presence of neighbors around the target residue but does not take their orientation and actual position into account. As a result, some different methods should be used to process the local steric context.

3. Atom grid analysis
To consider the steric context of the target amino acid residue, another approach was proposed. In it, we build a uniform 3D grid to which atoms of neighbor residues are projected. Probabilities of each rotamer are returned as an answer like in the first algorithm.

Grid creation includes the following steps:
Table 1. LightGBM accuracies for $\chi$ angles prediction

| Amino Acid | $\chi_1$ | $\chi_2$ | $\chi_3$ | $\chi_4$ | $\chi_5$ |
|------------|----------|----------|----------|----------|----------|
| ARG        | 0.66     | 0.72     | 0.52     | 0.55     | 0.94     |
| LYS        | 0.66     | 0.69     | 0.65     | 0.61     |          |
| GLN        | 0.67     | 0.65     | 0.60     |          |          |
| GLU        | 0.63     | 0.67     | 0.45     |          |          |
| MET        | 0.65     | 0.64     | 0.49     |          |          |
| HIS        | 0.68     | 0.53     |          |          |          |
| ILE        | 0.85     | 0.73     |          |          |          |
| LEU        | 0.78     | 0.80     |          |          |          |
| ASP        | 0.72     | 0.61     |          |          |          |
| ASN        | 0.70     | 0.70     |          |          |          |
| PHE        | 0.82     | 0.72     |          |          |          |
| PRO        | 0.82     | 0.98     |          |          |          |
| TRP        | 0.76     | 0.70     |          |          |          |
| TYR        | 0.81     | 0.71     |          |          |          |
| CYL        | 0.69     |          |          |          |          |
| SER        | 0.66     |          |          |          |          |
| THR        | 0.81     |          |          |          |          |
| VAL        | 0.81     |          |          |          |          |

(i) First, the grid is initialized with zeros; a new basis for the grid is selected. Here, $C_\beta$ atom of the target residue is located in the center of the grid.

(ii) For each neighbor, coordinates of its $C_\alpha$ and $C_\beta$ atoms are computed.

(iii) Let us denote local coordinates of an atom as $C$. Then, $\forall i,j,k : d_{i,j,k} \leq INTERACTION\_DIST \cdot G_{i,j,k} = G_{i,j,k} - INTERACTION\_DIST \cdot d_{i,j,k}$, where $d_{i,j,k}$ is the distance between the center of cell $(i,j,k)$ and $C$; $G_{i,j,k}$ is the value of the grid cell with coordinates $(u,j,k)$. This procedure is performed for every neighbor atom.

Finally, we get a 3D matrix with non-positive values. In it, zero values represent areas where the target side-chain can be located, where other directions are ‘penalized’ with negative values. Moreover, penalty values increase when moving closer to the center of a neighbor atom. $INTERACTION\_DIST$ value is selected equal to 4Å which is the minimum point for $C - C$ (carbon-carbon) van der Waals interaction.

The created grids are processed with Convolutional Neural Networks (CNN). In addition, some meta-parameters are appended to the output of Convolutional layers. These are the prior probabilities and features of the target amino acid. The latter is necessary as we do not distinguish between different amino acid types; instead, all samples are processed with the same methodology. Finally, the merged vector is processed with Dense layers.

Unfortunately, this approach did not show any significant improvement. Results of training neural networks of different structures (numbered 0 – 7) are presented in Figure 4. It can be seen that most of the models are badly overfit. This result may be explained by the fact that we omit the features of neighbor residues. In other words, it is impossible to discover to which amino acid type a neighbor atom corresponds.

Also, changing the value of $INTERACTION\_DIST$ may help. Usage of 4Å may be too strict, so this value may be changed to $C$ (carbon) van der Waals radius which equal to 1.7Å.
4. Protein Chain Segment Analysis

As it has been noticed, ‘steric clashes’ may happen not only between a side-chain and backbone but also between two side-chains. Unfortunately, both described approaches are unable to consider this issue.

A possible way to resolve all issues associated with clashes is predicting $\chi$ angle values not for a single residue but for all residues in a chain segment. It is proposed to collect samples of length equal to 64. This length was selected by analyzing the distribution of numbers of residues between the farthest neighbors along the chain.

Each sample is a matrix of shape $(64 \times M)$, where $M$ is the total number of one-hot-encoded features collected for each residue in the sample. The output is a matrix of shape $(64 \times 36)$. For each residue in a segment, we divide the $(-\pi, \pi)$ range into 36 parts. In the output matrix, each $i^{th}$ line contains 1 at position $j$ if the angle value is located in $j^{th}$ bin and 0 otherwise. These samples are processed with LSTM (Long Short-Term Memory) layers.

For evaluating the model, the following metric was proposed. For a residue $i$, let us denote the bin with the maximum predicted probability as $p_i$ and the original bin as $r_i$. Then:

$$m_i = \begin{cases} 
1, & \text{if } |p_i - r_i| \leq 3 \\
0, & \text{otherwise}
\end{cases}$$

Formula (1) implies that $m_i$ is equal to 1 if the most probable bin is no more than 3 bins away from the original bin; 0 otherwise. Having that, the whole sample is evaluated in the following way:

$$M(\text{pred}, \text{orig}) = \frac{1}{64} \sum_{i=1}^{64} m_i$$

The metric was applied to all predictions for the test samples. Figure 5 presents a histogram of these values. Currently, the method manages to predict about a half of a sample in the majority of cases.

Currently, this approach is developed only for $\chi_1$, so the next angles are supposed to be considered later. Another issue is that the proposed metric does not take the possibility of overlapping predictions into account. It means that one residue may be predicted several times in different samples, and these predictions should be merged.
5. Discussion
In the paper, several algorithms for rapid protein side-chain conformation prediction are described. In all of them, we try to consider the local context of the target residue to avoid steric clashes.

Three proposed algorithms use different approaches to including features of neighbor residues. Segment analysis algorithm aggregates the majority of the information about the steric context. At the same time, segment prediction requires resolving several significant issues. These include merging several predictions for a single residue and giving prediction for \( \chi_i, i > 1 \) angles.

We can conclude that side-chains of different residues should be predicted together to avoid ‘steric clashes’ with both backbone and other side-chains.

Acknowledgments
This work was partially supported by HSE University project ‘Development, research and application of optimization methods’.

References
[1] Levitt M 1976 *Journal of molecular biology* **104** 59–107
[2] Wilson C and Doniaich S 1989 *Proteins: Structure, Function, and Bioinformatics* **6** 193–209
[3] Dill K, Bromberg S, Yue K, Chan H, Ftebig K, Yee D and Thomas P 1995 *Protein science* **4** 561–602
[4] Bechini A 2013 *PloS one* **8** e59504
[5] Arab S, Sadeghi M, Eslahchi C, Pezeshk H and Sheari A 2010 *Bmc Bioinformatics* **11** 16
[6] Kini R and Evans H 1991 *Journal of biomolecular structure and dynamics* **9** 475–88
[7] MacKerell Jr, Alexander D, Banavali N and Folep N 2000 *Biopolymers: Original Research on Biomolecules* **56** 257–65
[8] Brooks S and Morgan B 1994 *Statistics and Computing* **4** 173–7
[9] Bower M, Cohen F and Dunbrack Jr R 1997 *Journal of molecular biology* **267** 1268–82
[10] Canutescu A, Shelenkov A and Dunbrack Jr R 2003 *Protein science* **12** 2001–14
[11] Bhuyan M and Gao X 2011 *BMC bioinformatics* **12** S10
[12] Lee C and Subbiah S 1991 *Journal of molecular biology* **217** 373–88
[13] Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P 2000 *Nucleic Acids Research* **28** 235–42
[14] Cock P, Antao T, Chang J, Chapman B, Cox C, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B et al. 2009 *Bioinformatics* **25** 1422–3
[15] Kyte J and Doolittle, R 1982 *Journal of molecular biology* **157** 105–32
[16] Taylor W 1986 *Journal of theoretical Biology* **119** 205–18
[17] Pommié C et al. 2004 *J. Mol. Recognit.* **17** 17–32
[18] Dunbrack Jr, Roland L and Karplus M 1993 *Journal of molecular biology* **230** 543–74