Abstract

Most computational methods used for predicting the 3D structure of proteins generate thousands of candidate structures, from which the "good" ones have to be selected. We present several scoring functions for discriminating correct from incorrect structures, based on the geometry of the amino acid (AA) interactions in proteins. Earlier, Krishnamoorthy and co-workers had developed a four-body contact scoring function. We represent proteins by the side chain centers of their AAs. Delaunay tessellation of this representation defines all sets of nearest neighbor quadruplets of AAs. Four-body contact scoring function (defined as log likelihood) is derived by the analysis of a diverse set of proteins with known structures. A test protein is characterized by the total score calculated as the sum of the individual log likelihoods of comprising amino acid quadruplets. We now propose three- and two-body scoring functions, thus developing a hierarchy of Delaunay-based scoring functions. The performance of the hierarchy follows the expected trend - four-body is better than three-body, which is better than the two-body function. In addition, the three-body definition discriminates triangles based on their buriedness; whether the triangles are on the surface or in the interior of the protein. The three-body scoring function with buriedness performs much better than the default three-body function. Finally, a combination of the three-body function with buriedness and the four-body scoring function proves to be the most accurate for decoy discrimination.

Proteins

Proteins are large biomolecules made of amino acids (AA) connected by a back-bone. There are 20 different AAs. Certain AAs tend to lie near certain others in 3D space, due to their chemical properties. Statistics of such AA contacts could be used to discriminate a correct structure from an incorrect one. The definition of a contact needs to be robust - if we use, say, 8 Å, then 7.9 Å is a contact, but 8.01 Å is not.

Voronoi and Delaunay tessellation

Given a set of points \( \{x, y, z\} \), the distance between any point \( x \) and \( y \) is \( d(x, y) = x - y \). The Voronoi cell of point \( x \) is \( V_x = \{ y \in \mathbb{R}^3 | d(x, y) \leq d(x, z) \} \). This could be thought of as a force around your house, defining the space that "belongs" to you. Delaunay tessellation is the dual graph of Voronoi diagram. A Delaunay edge defines two nearest neighbors - if you share a fence with someone, she is your neighbor [1, Chap. 1].

Tetrahedral Geometry of Proteins

Based on back-bone chain connectivity, there are five tetrahedra types. Taking these into account, we define the four-body scoring function as follows [3].

\[
Q_{ijk}^{b} = \log \frac{p_{ijk}}{f_{ijk}} = \log \left[ \frac{\text{observed frequency}}{\text{expected frequency}} \right]
\]

where (1)

\[
f_{ijk} = \frac{\text{number of type } i \text{ contacts with composition } (i,j,k)}{\text{total number of } i \text{ contacts}}
\]

and

\[
p_{ijk} = \frac{\text{number of amino acid of type } i \text{ in the data set}}{\text{total number of amino acids in the data set}}
\]

C = a number of amino acids of type \( i \) in the data set

\[
C = \frac{\text{number of amino acids of type } i \text{ in the data set}}{\text{total number of amino acids in the data set}}
\]

\[
Q_{ijk}^b = \frac{\text{log } Q_{ijk}^b}{\text{log } a_{ijk}}
\]

Three-Body Scoring Function

The motivation to investigate (3- and 3-body) contacts is to demonstrate that they are not as indicative of the correct structure as the four-body contacts. The Delaunay tetrahedra naturally defines triangles (for 3-body) and edges (for 2-body) for these contacts (without using arbitrary distance cut-offs). Similar to the five types of tetrahedra, there are three types of triangles based on back-bone chain connectivity.

Four-body outperforms all versions of 3-body when it comes to maximizing the discriminative power of the scoring function. The buriedness factor is not captured by the four-body scoring function. There are nine levels of buriedness.

Two-Body Scoring Function

To complete the hierarchy, we define log likelihood scores for 2-body contacts, which are defined by the edges of the Delaunay tetrahedra. There are two edge types based on connectivity - non-connected and connected (represented by \( \beta = 0 \) or \( \beta = 1 \)). There are 210 AA pairs, thus resulting in a total of 420 \( Q_{ijk}^b \) values.

Results

We tested the various scoring functions on the Decoys 31 Us data base [4], which has 151 protein sets with the correct structure and varying number of incorrect structures (decays). These decays were created by ten different computational methods. The quality of the decays is measured by its root mean square deviation from the native structure (RMSD). The correct structure has zero RMSD. The main goal of the scoring function is to make the correct structure as the top one among all structures for each decoy set. If it is not the top structure, the native should be ranked as high as possible.

Table 1: Comparison of native ranks assigned by various 2-, 3-, and 4-body scoring functions for the 151 decoy sets scored. The table gives the number of decay sets (out of 151) for which the native was the top, in the top 5%, etc.

We also calculated the native Z-Score and Spearman rank correlation coefficients for all the decays. The comparisons were similar to that of Table 1. Detailed results are given in our paper [2]. In summary:

1. Four-body without buriedness actually performs better than 3-body with buriedness, when all 9 buriedness classes are added;
2. Whens the \( \beta \) combinations are considered for 3-body, the combination 11101100 performs well, outperforming 3-body without buriedness;
3. 4-body outperforms all versions of 3-body when it comes to maximum first counts, and
4. 4-body + 3-body matches 4-body in performance, but some combinations do better than 4-body alone, in particular the combination 11111001-1-4-body.

References