Geometry-based Scoring Functions for Protein Fold Recognition

Andrew Fowler¹, Bala Krishnamoorthy, and Kelly Stratton²
Department of Mathematics, WSU

1. graduate student, 2. undergraduate student

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 composing amino acid quadruplets. We note properties of three- and two-body scoring functions. This discovers a hierarchy of Delaunay-based scoring functions. The performance of the hierarchy follows the expected trend—three-body is better than two-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.

1. Proteins

Proteins are large biomolecules made of amino acids (AA) connected by a backbone. There are 20 different AAs. Certain AAs tend to form a specific contact to one another, i.e., the definition is not based on arbitrary distance cut-offs. At the same time, the full Delaunay tessellation is a convex object, and contains a lot of big triangles, which do not represent meaningful contacts. Hence, we use a fairly large cut-off value of 10 Å to discard such spurious contacts.

2. Voronoi and Delaunay tessellation

Figure 1: All-atom model and backbone trace of a protein.

Figure 2: Voronoi diagram (red) and Delaunay triangulation (blue).

Given a set of points \( S = \{ x_1, x_2, \ldots, x_n \} \), the distance between any point \( x \) and \( x_i \) is \( d(x) = \| x - x_i \| \). The Voronoi cell of point \( x_i \) is \( \{ x \in \mathbb{R}^d | d(x) \leq d(x_i) \} \). This could be thought of as a fence 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].

3. Tetrahedral Geometry of Proteins

Figure 3: Five classes of Delaunay tetrahedra.

Tetrahedral Geometry of Proteins

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

\[ Q_{ijkl}^\text{d} = \log \frac{f_{ijkl}}{\binom{4}{i} \binom{4}{j} \binom{4}{k} \binom{4}{l}} \]

where

\[ f_{ijkl} = \text{number of type } ijkl \text{ quadruplets with composition } (ijkl) \]

\[ \binom{n}{k} = \frac{n!}{k!(n-k)!} \]

is the binomial coefficient.

4. Three-Body Scoring Function

The motivation to investigate (2- and) 3-body contacts is to demonstrate that they are not as indicative of the correct structure of the four-body contacts. The Delaunay tetrahedra naturally define 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 backbone chain connectivity.

Three-Body Scoring Function

\[ Q_{ijk} = \log \frac{f_{ijk}}{\binom{4}{i} \binom{4}{j} \binom{4}{k}} \]

where

\[ f_{ijk} = \text{number of type } ijk \text{ triangles in the data set} \]

\[ \binom{n}{k} = \frac{n!}{k!(n-k)!} \]

is the binomial coefficient.

5. Results

We tested the various scoring functions on the Decoys ‘R Us data base [4], which has 151 protein sets with the correct structure and varying number of incorrect structures (decoys). These decoys were created by ten different computational methods. The quality of the decoys measured by its root mean squared deviation from the native structure (RMSD). The correct structure has zero RMSD. The main goal of the scoring function is to rank 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 decoy sets (out of 151) for which the native was the top, the top 5, etc.

<table>
<thead>
<tr>
<th>Scoring Function</th>
<th>Native rank</th>
<th>Top 5</th>
<th>Top 10</th>
<th>Top 20</th>
<th>Top 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>No scoring</td>
<td>140</td>
<td>89</td>
<td>53</td>
<td>31</td>
<td>16</td>
</tr>
<tr>
<td>2-body (no buried)</td>
<td>14</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>3-body</td>
<td>14</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>4-body</td>
<td>14</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>3 + 4-body</td>
<td>14</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>3 + 4-body</td>
<td>14</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>14</td>
</tr>
</tbody>
</table>

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

• 3-body without buriedness actually performs better than 5-body with buriedness, when all 9 buriedness classes are added.
• When the \( z^2 \) combinations are considered for 4-body, the combination 111100 performs well, outperforming 3-body no-buriedness.
• 4-body outperforms versions of 3-body when it comes to maximize first counts and
• 4-body + 3-body matches 4-body in performance, but some combinations do better than 4-body alone, in particular the combination 11111001 (4-body).

References