Protein Structure Analysis

Iosif Vaisman

2015

Structure verification and validation

Anolea
Verify3D
Procheck
WhatIf

Bond lengths (Procheck)

<table>
<thead>
<tr>
<th>Bond</th>
<th>Labeling</th>
<th>Value</th>
<th>sigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-N</td>
<td>C-NH1</td>
<td>1.329</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>C-N</td>
<td>1.341</td>
<td>0.016</td>
</tr>
<tr>
<td>C-O</td>
<td>C-O</td>
<td>1.231</td>
<td>0.020</td>
</tr>
<tr>
<td>Calpha-C</td>
<td>CH1E-C</td>
<td>1.525</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>CH2G*-C</td>
<td>1.516</td>
<td>0.018</td>
</tr>
<tr>
<td>Calpha-Cbeta</td>
<td>CH1E-CH3E</td>
<td>1.521</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>CH1E-CH1E</td>
<td>1.540</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>CH1E-CH2E</td>
<td>1.530</td>
<td>0.020</td>
</tr>
<tr>
<td>N-Calpha</td>
<td>NH1-CH1E</td>
<td>1.458</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>NH1-CH2G*</td>
<td>1.451</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>N-CH1E</td>
<td>1.466</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Bond angles (Procheck)

<table>
<thead>
<tr>
<th>Angle</th>
<th>Labeling</th>
<th>Value</th>
<th>sigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-N-Calpha</td>
<td>C-NH1-CH1E</td>
<td>121.7</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>C-NH1-CH2G*</td>
<td>120.6</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>C-N-CH1E</td>
<td>122.6</td>
<td>5.0</td>
</tr>
<tr>
<td>Calpha-C-N</td>
<td>CH1E-CH3E</td>
<td>116.2</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>CH1E-CH1E</td>
<td>116.4</td>
<td>2.1</td>
</tr>
<tr>
<td>Calpha-C-O</td>
<td>CH1E-CH1E</td>
<td>116.9</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>CH1E-CH2E</td>
<td>120.8</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>CH2G*-C-O</td>
<td>120.8</td>
<td>2.1</td>
</tr>
</tbody>
</table>

Procheck output

a. Ramachandran plot quality - percentage of the protein's residues that are in the core regions of the Ramachandran plot.
b. Peptide bond planarity - standard deviation of the protein structure's omega torsion angles.
c. Bad non-bonded interactions - number of bad contacts per 100 residues.
d. Ca tetrahedral distortion - standard deviation of the ζ torsion angle (Ca, N, C, and Cβ).
e. Main-chain hydrogen bond energy - standard deviation of the hydrogen bond energies for main-chain hydrogen bonds.
f. Overall G-factor - average of different G-factors for each residue in the structure.
Procheck output - backbone G factors

Procheck output - all atom G factors

Anolea

High energy zones

Non-local contact maps

Melo et al., 1997

Model Quality Assessment

L.J. McGuffin et al., 2013
Local structure comparison algorithms

- Protein Local Structural Correlation problem
- Pattern Matching
- Pattern Discovery
- Pair-wise comparison
- Many-way comparison (n ≥ 3)

Root mean square deviation

Coordinate based RMSD

$$\text{RMSD}_{C}(A, B) = \min_{T} \sqrt{\frac{1}{m} \sum_{i=1}^{m} (A_i - TB_i)^2}$$

Distance based RMSD

$$\text{RMSD}_{D}(A, B) = \frac{1}{m} \left( \sum_{i=1}^{m} \sum_{j=1}^{\infty} (d_i^j - d_q^j)^2 \right)^{1/2}$$

Geometric Hashing

- Finding the maximum coincidence set is an NP-hard problem

Difference Distance Matrix Plot (DDMP)

Geometric Hashing

Reference frames

- Two points (basis pair) define a reference frame
- The coordinates of all points are computed in the reference frame (reference frame system)
- There will be pairs of points (from M and Q) with the same coordinates
- The number of such pairs depends on selection of reference frame and reference frame system resolution

Geometric Hashing Algorithm

Preprocessing
Hash table H is created. It has a bin for each cell in the frame systems. The coordinates of all points in each model frame system are calculated. If there is a point in the cell (p,q) in the frame system with basis (a_i,a_k), then (a_i,a_k) is placed in the bin H(p,q).

Recognition
A pair of points in the query is chosen as basis, and the coordinates of the other points are calculated. These coordinates are used as indices for H, and for each cell being indexed, a vote is given for the (model) basis pairs in the cell. The number of votes for a model basis pair is the number of coinciding points to the query (using the specified query basis pair).
Combinatorial Extension (CE) Algorithm

Structure alignment of phyocyanin (1CPC:L) to colicin A (1COL:A)

The solid line represents the optimal path built from AFPs. The dotted line represents the search area at every step of path extension.

The thick solid line represents alignment overlap both before and after optimization.

Calculation of distance

$D_{ij}$ for two AFPs $i$ and $j$ from the path

$D_{i}$ for single AFP $i$ from the path.

jCE and jFATCAT tool at RCSB

Pairwise structure alignment and pre-calculated 3D structure comparisons for the entire PDB

DaliLight algorithm for protein structure comparison

Search for common substructures by clique detection