Contacts-based indexing, retrieval and alignment of protein structures

Ahmet Sacan, I. Hakki Toroslu, and Hakan Ferhatosmanoglu
The Ohio State University

Introduction
- Protein structure is more conserved than sequence and can provide valuable information about functional and evolutionary relationships.
- The rapid growth in the protein structure database (PDB) makes finding similar protein structures a real challenge.

Methods
1. Residue Contacts
   - Observation: Residues in similar structures share similar inter-residue contacts.
   - Use Delaunay Tessellation to extract the contacts
   - Well-defined
   - Captures local geometry

2. Contact Strings
   - Encode the contacts of a residue based on their sequence order along the backbone.
   - Record both the amino acid (AA) type and the secondary structure (SS) information in the contact string.

3. Comparing Contact Strings
   - Sequence alignment can be used to obtain the distance between two contact strings.

4. Indexing
   - Seller’s Theorem (1974)
     - If a metric substitution matrix is used, the resulting alignment scores also form a metric.
   - We prove that:
     - "Weighted summation of two metric functions is also metric.
   - The metricity of the distances allows the use of distance-based indexing.

Experiments
Alignment Quality
- "ten difficult pairs" dataset. Our method is called "Vorometric"

<table>
<thead>
<tr>
<th>method</th>
<th>RMSD (Å)</th>
<th>%N (query)</th>
<th>quality (TM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>3.17</td>
<td>83.4</td>
<td>0.60</td>
</tr>
<tr>
<td>SSAP</td>
<td>4.37</td>
<td>88.1</td>
<td>0.59</td>
</tr>
<tr>
<td>DaliLite</td>
<td>2.82</td>
<td>80.0</td>
<td>0.61</td>
</tr>
<tr>
<td>Versicolor</td>
<td>2.28</td>
<td>51.7</td>
<td>0.56</td>
</tr>
<tr>
<td>Vorometric</td>
<td>3.02</td>
<td>84.8</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Similarity Search
- ASTRAL-90 database: 34,055 proteins

Classification
- ASTRAL-25: v1.65-v1.67 difference set

<table>
<thead>
<tr>
<th>Family</th>
<th>Superfamily</th>
<th>Fold</th>
<th>TM</th>
<th>%N</th>
<th>rmsd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vorometric-TM</td>
<td>96.7</td>
<td>94.9</td>
<td>97.6</td>
<td>0.74</td>
<td>87.2</td>
</tr>
<tr>
<td>Vorometric-cav</td>
<td>85.9</td>
<td>91.2</td>
<td>97.0</td>
<td>0.74</td>
<td>76.3</td>
</tr>
<tr>
<td>CE</td>
<td>84.6</td>
<td>91.9</td>
<td>94.1</td>
<td>0.77</td>
<td>78.2</td>
</tr>
<tr>
<td>SSEA</td>
<td>68.6</td>
<td>68.9</td>
<td>78.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLAST</td>
<td>48.9</td>
<td>52.5</td>
<td>52.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Structure similarity
- The similarity of two protein structures are defined in terms of their alignment.
- The quality of an alignment is measured by the following metrics:
  - Error: Root Mean Square Deviation
    \[ RMSD = \sqrt{\frac{\sum d_i^2}{N}} \]
  - Coverage: \( \%N \): percentage of residues aligned
  - Quality: a combined, normalized measure.
    \[ TM-score = \frac{1}{k_{target}} \sum_{r} \frac{1}{1 + (\frac{d_i}{k_{start}})^2} \]

Reference