Using PERL in Projects Related to Protein Structure Prediction

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins University

Email: ingo@jhu.edu

http://biostat.jhsph.edu/~iruczins

Collaborators

David Baker
University of Washington

Richard Bonneau
Institute for Systems Biology

Chris Bystroff
Rensselaer Polytechnic Institute

Charles Kooperberg
Fred Hutchinson Cancer Research Ctr

Carol Rohl
UC Santa Cruz

Kim Simons
Harvard University

Charlie Strauss
Los Alamos National Laboratory

Jerry Tsai
Texas A&M
What are Proteins?

The building blocks of proteins are amino acids.

2D and 3D Protein Structure

Both figures show the same protein, highlighting the tertiary and secondary structure.
Motivation

- What are proteins? Why do we care about them?
- Why do we care about protein structure?
- Why do we need to predict protein structures?
- How does the computational approach work?

Energy Landscape

The free energy of a structure changes with its geometry.
A Scoring Function for Ab Initio Protein Folding

\[ P(\text{structure}|\text{sequence}) \propto P(\text{sequence}|\text{structure}) \times P(\text{structure}) \]

**Sequence dependent:**
- hydrophobic burial
- residue pair interaction

**Sequence independent:**
- helix-strand packing
- strand-strand packing
- sheet configurations
- \(vdW\) interactions

**Hydrophobic Burial**

<table>
<thead>
<tr>
<th>amino acid</th>
<th>number of neighbours</th>
<th>6 8 10 12 14 16 18 20 22 24 26 28 30 31</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>7.8 7.1 7.2 7.3 7 6.5 6.8 7.3 8.7 9.5 10.6 12.4 14.3 17.8</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>0.4 0.3 0.3 0.5 0.7 1.1 1.6 2.2 2.5 3 3 3 3 3</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>7.9 9.4 9.5 9.6 8.6 7.6 5.8 4.7 3.9 2.7 2.5 2.7 2.6 3.2</td>
</tr>
<tr>
<td>E</td>
<td></td>
<td>8.7 9.3 10.1 10.3 9.8 8.3 6.4 4.9 3.2 2.4 1.8 1.8 1.5 1.1</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td>2.1 1.3 1.4 1.6 1.9 2.7 3.9 5.1 6.2 6.9 6.7 5.8 4.9 3.6</td>
</tr>
<tr>
<td>G</td>
<td></td>
<td>10.8 16.4 12.5 9.1 7.4 6.3 6 6 5.8 6.7 7.6 9.1 10.9 15.6</td>
</tr>
<tr>
<td>H</td>
<td></td>
<td>1.7 1.7 2 1.9 2.3 2.4 2.7 2.6 2.4 2 2.1 1.8 1.7 1.8</td>
</tr>
<tr>
<td>I</td>
<td></td>
<td>2.3 1.9 1.9 2.2 2.6 3.6 5.1 6.4 8 8.9 9.9 9.8 8.7 7.8</td>
</tr>
<tr>
<td>K</td>
<td></td>
<td>7.2 7.9 9.9 9.6 8.9 7.4 5 3.5 1.8 1.3 0.9 0.7 0.5</td>
</tr>
<tr>
<td>L</td>
<td></td>
<td>4.8 2.4 2.7 3.2 4.2 6.5 8 10.3 12.6 13.9 13.2 11.7 10.2 6.7</td>
</tr>
<tr>
<td>M</td>
<td></td>
<td>29 1 1 1.1 1.4 1.7 2 2.3 2.6 3.1 3.1 3.1 2.2</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>6.5 6.8 6.8 6.5 6.1 5.6 5.2 4 3.7 2.8 2.6 2.7 2.4 2.1</td>
</tr>
<tr>
<td>P</td>
<td></td>
<td>4.2 4 4.7 5 5.4 5.1 4.4 3.6 2.8 2.1 1.7 1.7 1.6 1.3</td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td>3.9 4 4.3 5.3 6 6.5 6.2 5.5 4.3 3.1 2.5 1.8 1.8 1</td>
</tr>
<tr>
<td>R</td>
<td></td>
<td>8.3 8.8 7.6 6.1 5.5 4.9 4.2 4.1 3.4 3.2 3 3.2 3.3</td>
</tr>
<tr>
<td>S</td>
<td></td>
<td>8.5 7.9 8.5 8.3 7.5 6.4 5.6 5.4 4.8 4.2 4.4 4.9 6.1 7.4</td>
</tr>
<tr>
<td>T</td>
<td></td>
<td>5.7 5.4 6.1 6.7 7.2 7 6.6 5.6 4.8 4.3 4.8 5 6.1 5.9</td>
</tr>
<tr>
<td>V</td>
<td></td>
<td>4 25 2.4 3.3 4.2 4.8 6.4 7.7 8.9 10.9 12.1 12.3 12.2 12.3</td>
</tr>
<tr>
<td>W</td>
<td></td>
<td>0 6 0.5 0.5 0.6 0.7 1.2 1.4 2 2.5 2.5 2.2 1.9 1.5 0.7</td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td>1.6 1.4 1.4 1.6 1.9 3 4.3 5.4 5.7 5.9 4.9 3.8 3.4 2.6</td>
</tr>
</tbody>
</table>
Energy Landscape (2)

True landscape  Our scoring function

Decoys tend to cluster near low energy states
Perl Scripts

Perl scripts are used all over the place! Some applications include:

- Accessing and transforming sequence data from local and remote databases
- Comparing amino acid sequences
- Extracting information from sequence alignments
- Comparing protein structures
- Parsing structures (sequences) into domains
- Creating decoy files and managing them
- …

Check out the bioperl tutorial (http://bio.perl.org/).

Beta-Sheet Motifs

Two possible motifs for 4-stranded sheets.
Local structures are easier to generate than non-local structures.

Likely Sheet Topologies

<table>
<thead>
<tr>
<th>L 1</th>
<th>L 2</th>
<th>L 3</th>
<th>L 4</th>
<th>L 5</th>
<th>L 6</th>
<th>L 7</th>
<th>L 8</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Diagram" /></td>
<td><img src="image2.png" alt="Diagram" /></td>
<td><img src="image3.png" alt="Diagram" /></td>
<td><img src="image4.png" alt="Diagram" /></td>
<td><img src="image5.png" alt="Diagram" /></td>
<td><img src="image6.png" alt="Diagram" /></td>
<td><img src="image7.png" alt="Diagram" /></td>
<td><img src="image8.png" alt="Diagram" /></td>
</tr>
</tbody>
</table>
References

Publications:


References

Web Pages:

- My Home Page http://www.biostat.jhsph.edu/~iruczins/
- Class Notes and Stuff http://www.biostat.jhsph.edu/~iruczins/teaching/misc/misc.html
- The Protein Data Bank http://www.rcsb.org/pdb/
- The Dunbrack Lab http://www.fccc.edu/research/labs/dunbrack/
- The Baker Lab http://depts.washington.edu/bakerpg/
- The CASP Page http://predictioncenter.llnl.gov/
- The Bioperl Page http://bioperl.org/
- The Beta Sheet Page http://www.biostat.jhsph.edu/~iruczins/sheets/sheets.html