Protein Structure Prediction and Analysis Tools

Jianlin Cheng, PhD
Assistant Professor
Department of Computer Science & Informatics Institute
University of Missouri, Columbia
2011
Sequence, Structure and Function

AGCWY......

Cell
Protein Folding Movie

http://www.youtube.com/watch?v=fvBO3TqJ6FE&feature=fvw
Alpha-Helix

Jurnak, 2003
Beta-Sheet

Anti-Parallel

Parallel
Non-Repetitive Secondary Structure

Beta-Turn

Loop
myoglobin

haemoglobin
Quaternary Structure: Complex

G-Protein Complex
Protein Structure Determination

- X-ray crystallography
- Nuclear Magnetic Resonance (NMR) Spectroscopy
- X-ray: any size, accurate (1-3 Angstrom (10^{-10} m)), sometime hard to grow crystal
- NMR: small to medium size, moderate accuracy, structure in solution
Pacific Northwest National Laboratory's high magnetic field (800 MHz, 18.8 T) NMR spectrometer being loaded with a sample.

Wikipedia, the free encyclopedia
Storage in Protein Data Bank

Search database
Search protein 1VJG
PDB Format (2C8Q, insulin)

HEADER           HORMONE                        06-DEC-05    2C8Q
TITLE  INSULINE (1SEC) AND UV LASER EXCITED FLUORESCENCE
COMPND          MOL_ID: 1;
COMPND          2 MOLECULE: INSULIN A CHAIN;
COMPND          3 CHAIN: A;
COMPND          4 MOL_ID: 2;
COMPND          5 MOLECULE: INSULIN B CHAIN;
COMPND          6 CHAIN: B
SOURCE          MOL_ID: 1;
SOURCE          2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE          3 ORGANISM_COMMON: HUMAN;
SOURCE          4 ORGAN: PANCREAS;
SOURCE          5 MOL_ID: 2;
SOURCE          6 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE          7 ORGANISM_COMMON: HUMAN;
SOURCE          8 ORGAN: PÁNCREAS
KEYWDS          LASER, UV, CARBOHYDRATE METABOLISM, HORMONE, DIABETES
KEYWDS          2 MELLITUS, GLUCOSE METABOLISM
EXPDTA          X-RAY DIFFRACTION
AUTHOR          X.VERNEDE, B.LAVALT, J.OHANA, D.NURIZZO, J.JOLY, L.JACQUAMET,
AUTHOR          2 F.FELISAZ, F.CIPRIANI, D.BOURGEOS
REVDAT          1 08-MAR-06    2C8Q    0
JRNL            AUTH  X.VERNEDE, B.LAVALT, J.OHANA, D.NURIZZO, J.JOLY,
JRNL            AUTH 2 L.JACQUAMET, F.FELISAZ, F.CIPRIANI, D.BOURGEOS
JRNL            TITL  UV LASER-EXCITED FLUORESCENCE AS A TOOL FOR THE
JRNL            TITL 2 VISUALIZATION OF PROTEIN CRYSTALS MOUNTED IN
JRNL            TITL 3 LOOPS.
JRNL            REF  ACTA CRYSTALLOGR., SECT. D V. 62 253 2006
JRNL            REFN  ASTM ABCRE6  DK ISSN 0907-4449
REMARK          2
REMARK          2 RESOLUTION. 1.95 ANGSTROMS.
REMARK          3
REMARK          3 REFINEMENT.
REMARK          3 PROGRAM : REFMAC 5.2.0005
REMARK          3 AUTHORS : MURSHUDOV, VAGIN, DODSON
REMARK          3
REMARK          3 REFINEMENT TARGET : MAXIMUM LIKELIHOOD
SEQRES 1 A 21 GLY ILE VAL GLU GLN CYS CYS THR SER ILE CYS SER LEU
SEQRES 2 A 21 TYR GLN LEU GLU ASN TYR CYS ASN
SEQRES 1 B 29 PHE VAL ASN GLN HIS LEU CYS GLY SER HIS LEU VAL GLU
SEQRES 2 B 29 ALA LEU TYR LEU VAL CYS GLY GLU ARG GLY PHE PHE TYR
SEQRES 3 B 29 THR PRO LYS
FORMUL 3 HOH *31(H2 O1)
HELIX 1 1 GLY A 1 CYS A 7 1 7
HELIX 2 2 SER A 12 ASN A 18 1 7
HELIX 3 3 GLY B 8 GLY B 20 1 13
HELIX 4 4 GLU B 21 GLY B 23 5 3
SSBOND 1 CYS A 6 CYS A 11 1555 1555
SSBOND 2 CYS A 7 CYS B 7 1555 1555
SSBOND 3 CYS A 20 CYS B 19 1555 1555
CRYST1 78.608 78.608 78.608 90.00 90.00 90.00 I 21 3 24
ORIGX1 1.000000 0.000000 0.000000 0.000000
ORIGX2 0.000000 1.000000 0.000000 0.000000
ORIGX3 0.000000 0.000000 1.000000 0.000000
SCALE1 0.012721 0.000000 0.000000 0.000000
SCALE2 0.012721 0.000000 0.000000 0.000000
SCALE3 0.000000 0.000000 0.012721 0.000000
ATOM 1 N GLY A 1 45.324 26.807 11.863 1.00 24.82 N
ATOM 2 CA GLY A 1 45.123 27.787 12.967 1.00 24.93 C
ATOM 3 C GLY A 1 43.756 27.627 13.605 1.00 25.16 C
ATOM 4 O GLY A 1 43.107 26.591 13.438 1.00 25.00 O
ATOM 5 N ILE A 2 43.313 28.661 14.323 1.00 25.21 N
ATOM 6 CA ILE A 2 42.050 28.622 15.065 1.00 25.39 C
ATOM 7 C ILE A 2 40.818 28.303 14.200 1.00 25.69 C
ATOM 8 O ILE A 2 39.935 27.565 14.635 1.00 25.56 O
ATOM 9 CB ILE A 2 41.816 29.917 15.917 1.00 25.39 C
Structure Visualization

- Rasmol  
  (http://www.umass.edu/microbio/rasmol/getras.htm)
- MDL Chime (plug-in)  
  (http://www.mdl.com/products/framework/chime/)
- Protein Explorer  
  (http://molvis.sdsc.edu/protexpl/frntdoor.htm)
- Jmol: http://jmol.sourceforge.net/
- Pymol: http://pymol.sourceforge.net/
Rasmol (1VJG)
Structure Analysis

• Assign secondary structure for amino acids from 3D structure
• Generate solvent accessible area for amino acids from 3D structure
DSSP server:  http://bioweb.pasteur.fr/seqanal/interfaces/dssp-simple.html
DSSP download:  http://swift.cmbi.ru.nl/gv/dssp/

DSSP Code:
H = alpha helix
G = 3-helix (3/10 helix)
I = 5 helix (pi helix)
B = residue in isolated beta-bridge
E = extended strand, participates in beta ladder
T = hydrogen bonded turn
S = bend
Blank = loop
DSSP Web Service

**DSSP** : Definition of secondary structure of proteins given a set of 3D coordinates (W.Kabsch, C. Sander)

[Image of input form]

```
[Image of PDB File input]

1vji or you can instead enter a PDB id.
```

http://bioweb.pasteur.fr/seqanal/interfaces/dssp-simple.html
<table>
<thead>
<tr>
<th></th>
<th>RESIDUE</th>
<th>AA STRUCTURE</th>
<th>BP1</th>
<th>BP2</th>
<th>ACC</th>
<th>N-H--&gt;O</th>
<th>O--&gt;H-N</th>
<th>N-H--&gt;O</th>
<th>O--&gt;H-N</th>
<th>TCO</th>
<th>KAPPA</th>
<th>ALPHA</th>
<th>PHI</th>
<th>PSI</th>
<th>X-CA</th>
<th>Y-CA</th>
<th>Z-CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>179</td>
<td>2,0.0</td>
<td>0, 0.0</td>
<td>0, 0.0</td>
<td>0.000</td>
<td>360.0</td>
<td>360.0</td>
<td>360.0</td>
<td>125.7</td>
<td>-8.6</td>
<td>43.0</td>
<td>43.9</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>K</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>123</td>
<td>2,0.4</td>
<td>37,0.1</td>
<td>37,0.2</td>
<td>-0.235</td>
<td>360.0</td>
<td>-108.7</td>
<td>87.0</td>
<td>151.4</td>
<td>-7.5</td>
<td>41.4</td>
<td>40.6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>-</td>
<td>39</td>
<td>GA</td>
<td>75</td>
<td>37,2.5</td>
<td>1,0.0</td>
<td>2,0.3</td>
<td>-0.593</td>
<td>34.7</td>
<td>-132.0</td>
<td>72.2</td>
<td>128.3</td>
<td>-4.3</td>
<td>39.5</td>
<td>39.6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Q</td>
<td>+a</td>
<td>40</td>
<td>GA</td>
<td>91</td>
<td>69,0.6</td>
<td>35,0.2</td>
<td>2,0.4</td>
<td>-0.639</td>
<td>26.0</td>
<td>179.8</td>
<td>86.4</td>
<td>132.7</td>
<td>-2.0</td>
<td>41.5</td>
<td>37.4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>I</td>
<td>-ab</td>
<td>41</td>
<td>73A</td>
<td>3</td>
<td>37,2.9</td>
<td>-2,0.3</td>
<td>2,0.5</td>
<td>-0.991</td>
<td>13.3</td>
<td>-156.5</td>
<td>129.4</td>
<td>131.5</td>
<td>-0.7</td>
<td>39.9</td>
<td>34.2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>E</td>
<td>-ab</td>
<td>42</td>
<td>74A</td>
<td>48</td>
<td>67,2.8</td>
<td>69,1.7</td>
<td>2,0.4</td>
<td>-0.910</td>
<td>14.8</td>
<td>-173.2</td>
<td>105.2</td>
<td>126.8</td>
<td>1.6</td>
<td>41.6</td>
<td>31.8</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>I</td>
<td>-ab</td>
<td>43</td>
<td>75A</td>
<td>0</td>
<td>35,2.5</td>
<td>37,2.6</td>
<td>2,0.5</td>
<td>-0.983</td>
<td>11.9</td>
<td>-162.4</td>
<td>124.9</td>
<td>124.4</td>
<td>1.7</td>
<td>40.3</td>
<td>23.2</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>C</td>
<td>-ab</td>
<td>44</td>
<td>76A</td>
<td>0</td>
<td>67,2.3</td>
<td>69,2.6</td>
<td>2,0.6</td>
<td>-0.931</td>
<td>6.5</td>
<td>-159.9</td>
<td>100.8</td>
<td>130.8</td>
<td>5.9</td>
<td>41.2</td>
<td>25.3</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>F</td>
<td>-ab</td>
<td>45</td>
<td>77A</td>
<td>0</td>
<td>35,2.2</td>
<td>37,3.0</td>
<td>2,0.5</td>
<td>-0.955</td>
<td>13.2</td>
<td>-169.0</td>
<td>109.5</td>
<td>117.1</td>
<td>2.7</td>
<td>40.2</td>
<td>21.8</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>V</td>
<td>+ab</td>
<td>46</td>
<td>78A</td>
<td>0</td>
<td>67,3.1</td>
<td>69,2.2</td>
<td>2,0.6</td>
<td>-0.926</td>
<td>34.8</td>
<td>71.1</td>
<td>116.5</td>
<td>129.9</td>
<td>5.6</td>
<td>40.1</td>
<td>19.4</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>G</td>
<td>-ab</td>
<td>47</td>
<td>79A</td>
<td>0</td>
<td>35,0.9</td>
<td>37,1.9</td>
<td>2,0.5</td>
<td>69,0.2</td>
<td>-0.921</td>
<td>70.2</td>
<td>-50.2</td>
<td>169.0</td>
<td>-146.4</td>
<td>5.3</td>
<td>39.9</td>
<td>15.6</td>
</tr>
<tr>
<td>12</td>
<td>D</td>
<td>&gt;&gt; S-</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>67,0.8</td>
<td>4,2.2</td>
<td>-2,0.3</td>
<td>-0.023</td>
<td>78.2</td>
<td>-51.3</td>
<td>115.5</td>
<td>151.8</td>
<td>4.2</td>
<td>41.6</td>
<td>12.4</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>S</td>
<td>3=&gt;S+</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>35,0.3</td>
<td>5,1.7</td>
<td>1,0.2</td>
<td>4,1.5</td>
<td>0.803</td>
<td>130.2</td>
<td>57.8</td>
<td>67.3</td>
<td>28.8</td>
<td>1.2</td>
<td>43.5</td>
<td>11.1</td>
</tr>
<tr>
<td>14</td>
<td>F</td>
<td>3=&gt;5S+</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>2,0.2</td>
<td>12,0.5</td>
<td>1,0.2</td>
<td>-1,0.2</td>
<td>0.884</td>
<td>108.5</td>
<td>46.5</td>
<td>-68.2</td>
<td>-33.2</td>
<td>-1.2</td>
<td>40.8</td>
<td>12.2</td>
</tr>
<tr>
<td>15</td>
<td>H</td>
<td>&lt;455+</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-3,0.6</td>
<td>12,0.3</td>
<td>64,0.2</td>
<td>-2,0.2</td>
<td>0.900</td>
<td>111.1</td>
<td>52.2</td>
<td>-68.9</td>
<td>-41.4</td>
<td>-0.0</td>
<td>41.1</td>
<td>15.7</td>
</tr>
<tr>
<td>16</td>
<td>N</td>
<td>&lt;SS-</td>
<td>0</td>
<td>0</td>
<td>71</td>
<td>-4,2.2</td>
<td>-2,0.2</td>
<td>30,0.1</td>
<td>-1,0.2</td>
<td>0.774</td>
<td>110.8</td>
<td>-127.0</td>
<td>-62.6</td>
<td>-26.6</td>
<td>-0.3</td>
<td>45.0</td>
<td>15.4</td>
</tr>
<tr>
<td>17</td>
<td>T</td>
<td>&gt;&gt;5-</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>-4,1.5</td>
<td>3,2.2</td>
<td>-5,0.2</td>
<td>8,0.4</td>
<td>0.741</td>
<td>36.4</td>
<td>174.6</td>
<td>83.1</td>
<td>25.3</td>
<td>-3.9</td>
<td>44.5</td>
<td>14.2</td>
</tr>
<tr>
<td>18</td>
<td>A</td>
<td>&gt;&gt;T3 &lt; +</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>-5,1.7</td>
<td>-1,0.2</td>
<td>1,0.3</td>
<td>-2,0.0</td>
<td>-0.199</td>
<td>68.4</td>
<td>29.2</td>
<td>-54.0</td>
<td>135.4</td>
<td>-3.4</td>
<td>46.6</td>
<td>11.0</td>
</tr>
<tr>
<td>19</td>
<td>G</td>
<td>&gt;&gt;T3 S+</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td>1,0.3</td>
<td>159,0.1</td>
<td>162,0.2</td>
<td>0.121</td>
<td>86.2</td>
<td>120.8</td>
<td>94.7</td>
<td>-21.4</td>
<td>-6.7</td>
<td>47.0</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>A</td>
<td>&gt;&gt;X-</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>-3,2.2</td>
<td>3,1.2</td>
<td>160,0.2</td>
<td>-1,0.3</td>
<td>-0.706</td>
<td>48.9</td>
<td>160.5</td>
<td>-79.7</td>
<td>117.6</td>
<td>-0.9</td>
<td>46.3</td>
<td>12.4</td>
</tr>
<tr>
<td>21</td>
<td>P</td>
<td>&gt;&gt;T3 S+</td>
<td>0</td>
<td>0</td>
<td>91</td>
<td>0,0.0</td>
<td>-1,0.2</td>
<td>0,0.0</td>
<td>159,0.0</td>
<td>0.677</td>
<td>91.8</td>
<td>60.1</td>
<td>-70.9</td>
<td>-17.3</td>
<td>-10.9</td>
<td>50.1</td>
<td>12.6</td>
</tr>
<tr>
<td>22</td>
<td>E</td>
<td>&gt;&gt;T3 S-</td>
<td>0</td>
<td>0</td>
<td>119</td>
<td>-3,0.0</td>
<td>-2,0.1</td>
<td>3,0.0</td>
<td>158,0.0</td>
<td>0.426</td>
<td>105.0</td>
<td>132.3</td>
<td>-87.9</td>
<td>-3.3</td>
<td>-11.4</td>
<td>49.4</td>
<td>16.3</td>
</tr>
<tr>
<td>23</td>
<td>C</td>
<td>&gt;&gt;S-</td>
<td>0</td>
<td>0</td>
<td>112</td>
<td>-3,1.2</td>
<td>-5,0.1</td>
<td>-6,0.2</td>
<td>0.730</td>
<td>80.2</td>
<td>98.1</td>
<td>62.8</td>
<td>28.1</td>
<td>-7.6</td>
<td>49.4</td>
<td>16.9</td>
<td></td>
</tr>
</tbody>
</table>
Solvent Accessibility

Size of the area of an amino acid that is exposed to solvent (water).

Maximum solvent accessible area for each amino acid is its whole surface area.

Hydrophobic residues like to be Buried inside (interior).
Hydrophilic residues like to be exposed on the surface.
Structure Comparison (Alignment)

- Are the structures of two protein similar?
- Are the two structure models of the same protein similar?
- Different measures (RMSD, GDT-TS (Zemla et al., 1999), MaxSub (Siew et al., 2000), TM score (Zhang and Skolnick, 2005))
Useful Structure Alignment Tools

- **CE**
  (http://cl.sdsc.edu/)

- **DALI**
  (http://www.ebi.ac.uk/dali/)

- **TM-Align:**
  http://zhang.bioinformatics.ku.edu/TM-align/
CE CALCULATE TWO Calculate structural alignment for two polypeptide chains either from the PDB or uploaded by the user.

Specify two polypeptide chains and optionally the similarity level and use of sequence information and then press the "Calculate Alignment" button. Selecting the appropriate similarity level will provide help on that specific similarity level.

Select Similarity Level: Medium

Use Sequence Information (optional)

Chain 1:
- PDB: CHAIN A
- User File: /casp7301/foldpro1.pdb
- Chain ID: 
- Use Fragment From: 
- To: 
- Sequence numbering

Chain 2:
- PDB: CHAIN B
- User File: /casp7301/ROSETTA_TS1.pdb
- Chain ID: 
- Use Fragment From: 
- To: 
- Sequence numbering

USR1: (size=395) vs USR2: (size=395)
Structure Alignment

Rmsd = 2.4 Å
Z-Score = 6.6
Sequence identity = 42.8%
Aligned/gap positions = 332/105

Sequence alignment based on structure alignment: Position numbers according to sequence (starting from 1) and according to PDB are given as 5555/PPP5, 5555 - sequence, PPP5 - PDB.

USR1:

USR2:

USR1: 4/5  PPQIRIPATYLRG5SKGVFRLDLPE-------SCRVQGEARDLPNBMVIGSPFDVEAA
USR1: 6/7  QITIRIPATYLRG5SKGVFRL-------DLPLQSCRVQGEARDLPNBMVIGSPFDVEAA
USR1: 97/98  HDNGK5GATSS5SKCVILSG5QPHQTVVLGFQV1DKUFVDSG5RCNLSTGACAFAL
USR1: 57/58  HDNGK5GATSS5SKCVILSG5QPHQTVVLGFQV1DKUFVDSG5RCNLSTGACAFAL
USR1: 117/118  HAGLVDFARIFEDGICEVRIDQAN1GKTI1AMHV5GQQVQETGDF6LGVTFPARE1VL
USR1: 117/118  HAGLVDFARIFEDGICEVRIDQAN1GKTI1AMHV5GQQVQETGDF6LGVTFPARE1VL

USR2: 4/5  PPQIRIPATYLRG5SKGVFRLDLPE-------SCRVQGEARDLPNBMVIGSPFDVEAA
USR2: 6/7  QITIRIPATYLRG5SKGVFRL-------DLPLQSCRVQGEARDLPNBMVIGSPFDVEAA
USR2: 97/98  HDNGK5GATSS5SKCVILSG5QPHQTVVLGFQV1DKUFVDSG5RCNLSTGACAFAL
USR2: 57/58  HDNGK5GATSS5SKCVILSG5QPHQTVVLGFQV1DKUFVDSG5RCNLSTGACAFAL
USR2: 117/118  HAGLVDFARIFEDGICEVRIDQAN1GKTI1AMHV5GQQVQETGDF6LGVTFPARE1VL
USR2: 117/118  HAGLVDFARIFEDGICEVRIDQAN1GKTI1AMHV5GQQVQETGDF6LGVTFPARE1VL
Notation: protein structure 1D, 2D, 3D
Goal of structure prediction

• Epstein & Anfinsen, 1961: sequence uniquely determines structure

• INPUT: sequence
• OUTPUT: 3D structure and function
CASP – Olympics of Protein Structure Prediction

- Critical Assessment of Techniques of Protein Structure Prediction
- Blind Test, Independent Evaluation
- CASP9: 116 targets
1D Structure Prediction

• Secondary structure
• Solvent accessibility
• Disordered regions
• Domain boundary
1D: Secondary Structure Prediction

Coil

MWLKKFGINLLIGQSV...

CCCC

HHHHH

CCC

SSSSS

Neural Networks + Alignments

CCCCHHHHHHCCCSSSSS...

Cheng, Randall, Sweredoski, Baldi. *Nucleic Acid Research*, 2005
Widely Used Tools (~78-80%)

**SSpro 4.1**: http://sysbio.rnet.missouri.edu/multicom_toolbox/

**Distill**: http://distill.ucd.ie/porter/

**PSI-PRED**: http://bioinf.cs.ucl.ac.uk/psipred/psiform.html

  software is also available

**SAM**: http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html

**PHD**: http://www.predictprotein.org/
1D: Solvent Accessibility Prediction

Exposed

Buried

MWLKKFGINLLIGQSV...

Neural Networks + Alignments

eeeeeeebbbbbbbbbeeeeeebbb...

Accuracy: 79% at 25% threshold

Cheng, Randall, Sweredoski, Baldi. *Nucleic Acid Research*, 2005
Widely Used Tools (78%)

- ACCpro 4.1: software: http://sysbio.rnet.missouri.edu/multicom_toolbox/
- SCRATCH: http://scratch.proteomics.ics.uci.edu/
- PHD: http://www.predictprotein.org/
- Distill: http://distill.ucd.ie/porter/
1D: Disordered Region Prediction Using Neural Networks

Disordered Region

MWLKKFGINLLIGQSV...

1D-RNN

OOOOODDDDOOOOOO...

93% TP at 5% FP

Deng, Eickholt, Cheng. BMC Bioinformatics, 2009
Tools

PreDisorder: http://sysbio.rnet.missouri.edu/multicom_toolbox/

A collection of disorder predictors:
http://www.disprot.org/predictors.php

Deng, Eickholt, Cheng. BMC Bioinformatics, 2009
1D: Protein Domain Prediction Using Neural Networks

HIV capsid protein

Domain 1

Boundary

Domain 2

MWLKKFGINLLIGQSV… + SS and SA

1D-RNN

NNNNNNNBBBNNNNNN…

Inference/Cut

Domains

Top *ab-initio* domain predictor in CAFASP4

Tools

DOMAC: http://casp.rnet.missouri.net/domac.html

DOMAC: An Accurate, Hybrid Protein Domain Prediction Server

Reference:

Dr. Jianlin Cheng’s Bioinformatics and Systems Biology Laboratory
Department of Computer Science
University of Missouri

Cheng, Nucleic Acids Research, 2007
DoBo
Protein domain boundary prediction by integrating evolutionary signals and machine learning

Have a question? Maybe it’s answered in the FAQ

Web: http://sysbio.rnet.missouri.edu/multicom_toolbox/index.html

Reference:
1. Input query
   LNKGQRHICREIIMS...

2. Identify homologous sequences w/ PSI-BLAST
   nr protein database

3. Extract pairwise alignments
   - Query 1: LNKGQRHICREIIMSNDIETQDELVRLEAGFNVTQATVSRDIKEMQLVKVPMANGRY 60
   - Sbjct 1: MNKGQRHICREIIMANKEIETQDELVDLRSNEGFNTQTATVSRDIKELHLVKVPLHDG 60
   - ...
   - Query 6: RHICREIIMSNDIETQDELVRLEAGFNVTQATVSRDIKEMQLVKVPMANGRYKSYL 65
   - Sbjct 5: RHSKIEELINKYEETQEDLITEYREAGINTQTATVSRDIQMKLTVKMTKSGYKYYAAY 64
   - ...
   - Query 1: LNKGQRHICREIIMSNDIETQDELVRLEAGFNVTQATVSRDIKEMQLVKVPMANGRY 60
   - Sbjct 1: MNKGQRHICREIIMANKEIETQDELVDLRSNEGFNTQTATVSRDIKELHLVKVPLHDG 60

4. Form multiple sequence alignment

5. Identify domain boundary signals
   - Gap 45 residues or longer
   - Remaining sequence longer than 45 residues
   - Domain boundary signal (indicated by large arrows)
Project 1

- Predict Secondary Structure, Solvent Accessibility, Disorder Regions of soybean transcription factors
- Data: http://casp.rnet.missouri.edu/marc/muii_7005/SEQ_TFP_90_2500.txt
- Select 10 proteins to make predictions
2D: Contact Map Prediction

3D Structure

2D Contact Map

Distance Threshold = 8Å°

Cheng, Randall, Sweredoski, Baldi. *Nucleic Acid Research*, 2005
Contact Prediction

- SVMcon: http://casp.rnet.missouri.edu/svmcon.html
- NNcon: http://casp.rnet.missouri.edu/nncon.html
- SCRATCH: http://scratch.proteomics.ics.uci.edu/
- SAM: http://compbio.soe.ucsc.edu/HMM-apps/HMM-applications.html
NNcon: Protein Contact Map Prediction Using Artificial Neural Networks (Help)

Two Methodologies for 3D Structure Prediction

• AB Initio Method (physical-chemical principles / molecular dynamics, knowledge-based approaches)

• Template-Based Method (knowledge-based approaches)
Two Approaches

• **Ab Initio Structure Prediction**
  
  Physical force field – protein folding
  Contact map - reconstruction

• **Template-Based Structure Prediction**

Query protein

MWLKKFGINKH…

Protein Data Bank

MWLKKFGINLIGQSV…

Simulation

Select structure with minimum free energy

MWLKKFGININKH…

Fold

Recognition

Alignment

Template
Protein Energy Landscape

C. Park, 2005
Markov Chain Monte Carlo Simulation

Template-Based Structure Prediction

1. Template identification
2. Query-template alignment
3. Model generation
4. Model evaluation
5. Model refinement

Notes: if template is easy to identify, it is often called comparative Modeling or homology modeling. If template is hard to identify, it is often called fold recognition.
TARGET

ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCCLKIDD
VPERLIPERASFQWMNDK

TEMPLATE

ASILPKRLFGNCEQTSDEGLKIERTPPLVPHEISAQNVCCLKIDDVDPERLIP
MSVIPKRLYGNCETSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE

A. Fisher, 2005
Modeller

- Need an alignment file between query and template sequence in the PIR format
- Need the structure (atom coordinates) file of template protein
- You need to write a simple script (Python for version 8.2) to tell how to generate the model and where to find the alignment file and template structure file.
- Run Modeller on the script. Modeller will automatically copy coordinates and make necessary adjustments to generate a model.
- See project step 5-8 for more details.
An PIR Alignment Example

Template id
Template structure file id
Structure determination method
Start index
End index
Query sequence id

> P1; 1SDMA
structureX: 1SDMA: 1: : 344: : : :
KIRVYCRLRPLCEKEIIAKERNARIVSVDEFTVEHLLWDDAKQHMYDRVFDGNATOQQDDVFDKYL
VQSAVGDGYNCIFAYGQTGSGLKTFITLYGADSNPGLTPTRAMSELFRIMKKDSNKFSSLKAYMVELY
QDTLVDLLLKPQAKRLKDIKDKSKGVMVSENVTVVISTYEEKLTIIQRGSEQHRGTGGTGLMNEQS
SRSHLIVSVTIESTNLQTQAIARGKLSFVLADAGSERSVKEAQSIKSLSALGDIVISALSSGNQHIP
YRNHKLTMMSDSLGGNAKTLTMFVNISPAESNLDETHNSLTYASRVRSIVNDPSKNVSSKEVARLK
KLVSYWELEEIQDE*

> P1; bioinfo

NIRVIARVRPVTKEEGHPEATNAVTFDADDDSIHLLHKGKPVSFELDKVFSPQASQODVFEQEVQ
ALVTSCIDGFGVNCIFAYGQTGAGKTYMEGTAENPGINQRALQLLFSVQEKAADWETYTIIVSAAE
IYNEVLRDLGKEPEQKLEIRLCPDGSQLYPVGLTEFQVQSVDDINKVFEGHTNRTTEETFNLNE
HSSRSHALLIVTVRGVDCSTGLRTTGKLNLVLGDSERSVQKSGAEGSRLREAQHIINKSLSALGDVIAALRSRQGHVPFRNSKLYTLQDLSGDSKTLMVV-----QVSPVEKNTSETLYSLKFAER------VR*
| ATOM | 1  | N   | LYS | 1   | -3.978 | 26.298 | 113.043 | 1.00 | 31.75 | N     |
| ATOM | 2  | CA  | LYS | 1   | -4.532 | 25.067 | 113.678 | 1.00 | 31.58 | C     |
| ATOM | 3  | C   | LYS | 1   | -5.805 | 25.389 | 114.448 | 1.00 | 30.38 | C     |
| ATOM | 4  | O   | LYS | 1   | -6.887 | 24.945 | 114.072 | 1.00 | 32.68 | O     |
| ATOM | 5  | CB  | LYS | 1   | -3.507 | 24.446 | 114.631 | 1.00 | 34.97 | C     |
| ATOM | 6  | CG1 | VAL | 4   | -3.743 | 22.970 | 114.942 | 1.00 | 36.49 | C     |
| ATOM | 7  | CD  | LYS | 1   | -3.886 | 22.172 | 113.644 | 1.00 | 39.52 | C     |
| ATOM | 8  | CE  | LYS | 1   | -3.318 | 20.766 | 113.761 | 1.00 | 41.58 | C     |
| ATOM | 9  | NZ  | LYS | 1   | -1.817 | 20.761 | 113.756 | 1.00 | 43.48 | N     |
| ATOM | 10 | N   | ILE | 2   | -5.687 | 26.161 | 115.522 | 1.00 | 26.16 | N     |
| ATOM | 11 | CA  | ILE | 2   | -6.867 | 26.500 | 116.302 | 1.00 | 22.75 | C     |
| ATOM | 12 | C   | ILE | 2   | -7.887 | 27.226 | 115.439 | 1.00 | 21.35 | C     |
| ATOM | 13 | O   | ILE | 2   | -7.565 | 28.200 | 114.770 | 1.00 | 20.95 | O     |
| ATOM | 14 | CB  | ILE | 2   | -6.513 | 27.377 | 117.523 | 1.00 | 21.68 | C     |
| ATOM | 15 | CG1 | ILE | 2   | -5.701 | 26.563 | 118.526 | 1.00 | 21.13 | C     |
| ATOM | 16 | CG2 | ILE | 2   | -7.782 | 27.875 | 118.200 | 1.00 | 18.96 | C     |
| ATOM | 17 | CD1 | ILE | 2   | -5.368 | 27.325 | 119.787 | 1.00 | 21.39 | C     |
| ATOM | 18 | N   | ARG | 3   | -9.120 | 26.737 | 115.461 | 1.00 | 22.04 | N     |
| ATOM | 19 | CA  | ARG | 3   | -10.214| 27.327 | 114.693 | 1.00 | 23.95 | C     |
| ATOM | 20 | C   | ARG | 3   | -10.783| 28.563 | 115.400 | 1.00 | 22.82 | C     |
| ATOM | 21 | O   | ARG | 3   | -10.771| 28.645 | 116.629 | 1.00 | 22.62 | O     |
| ATOM | 22 | CB  | ARG | 3   | -11.327| 26.290 | 114.510 | 1.00 | 26.34 | C     |
| ATOM | 23 | CG  | ARG | 3   | -11.351 | 25.586 | 113.161 | 1.00 | 30.68 | C     |
| ATOM | 24 | CD  | ARG | 3   | -10.004 | 25.034 | 112.771 | 1.00 | 35.43 | C     |
| ATOM | 25 | NE  | ARG | 3   | -10.104 | 24.072 | 111.672 | 1.00 | 43.37 | N     |
| ATOM | 26 | CZ  | ARG | 3   | -10.575 | 24.350 | 110.458 | 1.00 | 46.04 | C     |
| ATOM | 27 | NH1 | ARG | 3   | -10.997 | 25.572 | 110.168 | 1.00 | 48.68 | N     |
| ATOM | 28 | NH2 | ARG | 3   | -10.627 | 23.400 | 109.532 | 1.00 | 48.37 | N     |
| ATOM | 29 | N   | VAL | 4   | -11.278 | 29.524 | 114.630 | 1.00 | 20.49 | N     |
| ATOM | 30 | CA  | VAL | 4   | -11.853 | 30.724 | 115.225 | 1.00 | 17.59 | C     |
| ATOM | 31 | C   | VAL | 4   | -13.082 | 31.211 | 114.471 | 1.00 | 18.31 | C     |
| ATOM | 32 | O   | VAL | 4   | -13.030 | 31.446 | 113.264 | 1.00 | 16.37 | O     |
| ATOM | 33 | CB  | VAL | 4   | -10.834 | 31.872 | 115.272 | 1.00 | 19.94 | C     |
| ATOM | 34 | CG1 | VAL | 4   | -11.512 | 33.168 | 115.759 | 1.00 | 15.64 | C     |
| ATOM | 35 | CG2 | VAL | 4   | -9.668  | 31.489 | 116.168 | 1.00 | 15.45 | C     |
# Homology modelling by the automodel class

from modeller.automodel import *  # Load the automodel class

log.verbose()  # request verbose output
env = environ()  # create a new MODELLER environment to build this model in

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

# PIR alignment file name
alnfile = 'bioinfo.pir',  # alignment filename

# Template structure file id
knowns = '1SDMA',  # codes of the templates

# Query sequence id
sequence = 'bioinfo')  # code of the target

a = automodel(env,
              alnfile = 'bioinfo.pir',  # alignment filename
              knowns = '1SDMA',  # codes of the templates
              sequence = 'bioinfo')  # code of the target

a.starting_model= 1  # index of the first model
a.ending_model  = 1  # index of the last model

# (determines how many models to calculate)
a.make()  # do the actual homology modelling

Where to find structure file
PIR alignment file name
Template structure file id
Query sequence id
Output Example

Command: mod8v2  bioinfo.py
Homology modelling for entire genomes

Number of ORFs
Number of ORFs with PDB hit

Organism

B. Rost, 2005
%Sequence Identity: percent of identical residues in alignment
RMSD: square root of average distance between predicted structure and native structure.
3D Structure Prediction Tools

- MULTICOM (http://sysbio.rnet.missouri.edu/multicom_toolbox/index.html)
- I-TASSER (http://zhang.bioinformatics.ku.edu/I-TASSER/)
- Robetta (http://robetta.bakerlab.org/)
- 3D-Jury (http://bioinfo.pl/Meta/)
- FFAS (http://ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl)
- Pcons (http://pcons.net/)
- Sparks (http://phyyz4.med.buffalo.edu/hzhou/anonymous-fold-sp3.html)
- FUGUE (http://www-cryst.bioc.cam.ac.uk/%7Efugue/prfsearch.html)
- FOLDpro (http://mine5.ics.uci.edu:1026/foldpro.html)
- SAM (http://www.cse.ucsc.edu/research/compbio/sam.html)
- Phyre (http://www.sbg.bio.ic.ac.uk/~phyre/)
- 3D-PSSM (http://www.sbg.bio.ic.ac.uk/3dpssm/)
- mGenThreader (http://bioinf.cs.ucl.ac.uk/psipred/psiform.html)
Protein Model Quality Assessment

APOLLO: assessing protein single or multiple model(s) (help)
Evaluating the absolute and/or relative qualities of multiple models or a single model

Upload a compressed file (i.e. zip or tar.gz) containing multiple models OR a single model text file in PDB format: (two multiple models examples: example.zip, example.tar.gz; a single model file example: example)

OR paste a single model in PDB format: (example)

(Optional) Email address: (where the evaluation results will be sent to)

http://sysbio.rnet.missouri.edu/apollo/
## APOLLO Output

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Absolute Score</th>
<th>Average Pairwise GDT-TS Score</th>
<th>Refined Average Pairwise Q Score</th>
<th>Local Quality (click to enlarge)</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUARK_TS1</td>
<td>0.713</td>
<td>0.619</td>
<td>0.654</td>
<td></td>
</tr>
<tr>
<td>BAKER-ROSETTASERVER_TS1</td>
<td>0.668</td>
<td>0.503</td>
<td>0.516</td>
<td></td>
</tr>
<tr>
<td>MULTICOM-NOVEL_TS1</td>
<td>0.649</td>
<td>0.638</td>
<td>0.811</td>
<td></td>
</tr>
</tbody>
</table>
Application of Structure Prediction

• Structure prediction is improving
• Template-based structure become more and more practical. Particularly, comparative / homology modeling is pretty accurate in many cases.
• Comparative modeling has been widely used in drug design.
• Protein structure prediction (both secondary and tertiary) has become an indispensable tool of investigating function of proteins and mechanisms of biological processes.
Baker and Sali (2000)

J. Pevsner, 2005
Project 2

• Select 5 soybean proteins
• Predict 3D structures
• Visualize the structures

(data: http://casp.rnet.missouri.edu/marc/muii_7005/SEQ_TFP_90_2500.txt)