Residue contact and β-sheet Protein Structures Prediction

Jérôme Waldispühl, PhD
School of Computer Science,
McGill Centre for Bioinformatics,
McGill University

http://csb.cs.mcgill.ca
Secondary structure prediction methods reached 77% per residue accuracy (E.g. PSI-pred), but performance are weaker on β-strands.

Why?

β-sheets are stabilized through long-range contacts. Other, elements (e.g. helices) may benefit of long-range contacts too.

Objective:

Prediction of residue contacts and super-secondary structure.
Protein Structure

Sequence

A_1  A_2  A_3  A_4  A_5

β-sheet

Long-range interactions stabilize β-sheet.

Source: http://www.cryst.bbk.ac.uk/PPS2/

Structure

α-helix

β-strand
Residue contacts
Contact maps
Interpreting contact maps
• **Profcon**: Neural Network Approach
• **BetaPro**: Hierarchical Using Neural Networks, Alignments, and Graph Algorithms
• **SMURF**: Markov Random Field
• **tFolder**: Abstract template & ensemble prediction
Combine secondary structure and solvent accessibility predictions, and multiple sequence alignment information.

• Local information from immediate residue environment,
• Local information from connecting segment,
• Global information.
For each residue i, j involved, we define a window of size 9 centered around them.

Each residue position is characterized by:

- 20 units for the evolutionary informations (frequency of the 20 obtained from a PSI-BLAST multiple sequence alignment at 80% of homology),
- 1 unit N- or C- terminal,
- 4 units for the secondary structure prediction (PROFphd),
- 3 units for the solvent accessibility prediction (PROFphd),
- 1 unit for the conservation weight.

**Total**: $18 = 522$ units + coarse grained biophysical classification of the contact (hydrophobic-hydrophobic, polar-polar, charged-polar, opposite charges, same charge, aromatic-aromatic, others)
Local information from connecting segment
Design a window for the five consecutive residue that spanned the interval.
Each residue position is characterized by the same informations as contact residues + length of the segment, amino acid and secondary structure composition, SEG-low-complexity.
Total: 180 units

Global information
23 units for amino acid and secondary structure composition + 4 units for the protein length.
## PROFcon Results

### Table 1. Benefit from using connecting segments

<table>
<thead>
<tr>
<th>Method</th>
<th>Sequence separation</th>
<th>Nprot</th>
<th>Acc</th>
<th>Err&lt;sub&gt;Acc&lt;/sub&gt;</th>
<th>Cov</th>
<th>Err&lt;sub&gt;Cov&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local only</td>
<td>6</td>
<td>633</td>
<td>29.7</td>
<td>0.5</td>
<td>8.6</td>
<td>0.1</td>
</tr>
<tr>
<td>PROFcon</td>
<td>6</td>
<td>633</td>
<td>32.4</td>
<td>0.5</td>
<td>9.8</td>
<td>0.2</td>
</tr>
<tr>
<td>Local only</td>
<td>24</td>
<td>621</td>
<td>19.5</td>
<td>0.5</td>
<td>8.8</td>
<td>0.3</td>
</tr>
<tr>
<td>PROFcon</td>
<td>24</td>
<td>621</td>
<td>20.0</td>
<td>0.5</td>
<td>9.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

### Table 2. Improvement through evolutionary information

<table>
<thead>
<tr>
<th>Nali&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Sequence separation</th>
<th>Nprot&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Acc&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Err&lt;sub&gt;Acc&lt;/sub&gt;&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Cov&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Err&lt;sub&gt;Cov&lt;/sub&gt;&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–14</td>
<td>6</td>
<td>138</td>
<td>23.0</td>
<td>1.0</td>
<td>9.5</td>
<td>0.8</td>
</tr>
<tr>
<td>15–49</td>
<td></td>
<td>123</td>
<td>31.0</td>
<td>1.0</td>
<td>9.8</td>
<td>0.5</td>
</tr>
<tr>
<td>50–199</td>
<td></td>
<td>187</td>
<td>35.6</td>
<td>0.9</td>
<td>9.7</td>
<td>0.3</td>
</tr>
<tr>
<td>≥200</td>
<td></td>
<td>185</td>
<td>37.0</td>
<td>0.9</td>
<td>10.2</td>
<td>0.4</td>
</tr>
<tr>
<td>0–14</td>
<td>24</td>
<td>132</td>
<td>13.2</td>
<td>0.7</td>
<td>10.0</td>
<td>1.0</td>
</tr>
<tr>
<td>15–49</td>
<td></td>
<td>120</td>
<td>19.0</td>
<td>1.0</td>
<td>9.1</td>
<td>0.7</td>
</tr>
<tr>
<td>50–199</td>
<td></td>
<td>185</td>
<td>21.5</td>
<td>0.9</td>
<td>9.0</td>
<td>0.4</td>
</tr>
<tr>
<td>≥200</td>
<td></td>
<td>184</td>
<td>24.2</td>
<td>0.9</td>
<td>9.5</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<sup>a</sup>Number of proteins in multiple sequence alignment used to extract evolutionary profiles.

<sup>b</sup>As in Table 1; note all values compiled for the first L/2 predictions (Methods).
PROFcon Results: Length & sequence separation

Table 3. Performance versus protein length

<table>
<thead>
<tr>
<th>$L^a$</th>
<th>Sequence separation$^b$</th>
<th>Nprot$^b$</th>
<th>Acc$^b$</th>
<th>Err$^b_{\text{Acc}}$</th>
<th>Cov$^b$</th>
<th>Err$^b_{\text{Cov}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤100</td>
<td>6</td>
<td>78</td>
<td>31.0</td>
<td>2.0</td>
<td>18.0</td>
<td>1.0</td>
</tr>
<tr>
<td>101–200</td>
<td>230</td>
<td>191</td>
<td>33.9</td>
<td>0.9</td>
<td>8.6</td>
<td>0.2</td>
</tr>
<tr>
<td>201–300</td>
<td>134</td>
<td>633</td>
<td>32.4</td>
<td>0.5</td>
<td>9.8</td>
<td>0.2</td>
</tr>
<tr>
<td>301–400</td>
<td>24</td>
<td>66</td>
<td>19.0</td>
<td>1.0</td>
<td>22.0</td>
<td>2.0</td>
</tr>
<tr>
<td>All</td>
<td>6</td>
<td>621</td>
<td>20.0</td>
<td>0.5</td>
<td>9.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Diagram: Accuracy vs. sequence separation for different L values (L/5, L/10, L/20, random).

Errors: $\text{Err}_{\text{Acc}}$, $\text{Err}_{\text{Cov}}$.
## PROFcon Results: Structural classes

Table 4. Performance differs between structural classes

<table>
<thead>
<tr>
<th>SCOP class</th>
<th>Sequence separation(^a)</th>
<th>Nprot(^a)</th>
<th>Acc(^a)</th>
<th>Err(_{Acc})^a</th>
<th>Imp(^a)</th>
<th>Err(_{Imp})^a</th>
<th>Cov(^a)</th>
<th>Err(_{Cov})^a</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-alpha</td>
<td>6</td>
<td>131</td>
<td>24.0</td>
<td>1.0</td>
<td>14.3</td>
<td>0.8</td>
<td>11.0</td>
<td>0.7</td>
</tr>
<tr>
<td>All-beta</td>
<td></td>
<td>103</td>
<td>35.0</td>
<td>1.0</td>
<td>13.0</td>
<td>0.7</td>
<td>7.8</td>
<td>0.3</td>
</tr>
<tr>
<td>Alpha + beta</td>
<td></td>
<td>119</td>
<td>36.0</td>
<td>1.0</td>
<td>14.8</td>
<td>0.6</td>
<td>10.0</td>
<td>0.4</td>
</tr>
<tr>
<td>Alpha/beta</td>
<td></td>
<td>169</td>
<td>35.9</td>
<td>0.8</td>
<td>19.4</td>
<td>0.5</td>
<td>8.7</td>
<td>0.2</td>
</tr>
<tr>
<td>All-alpha</td>
<td>24</td>
<td>128</td>
<td>13.5</td>
<td>0.6</td>
<td>9.4</td>
<td>0.6</td>
<td>9.9</td>
<td>0.8</td>
</tr>
<tr>
<td>All-beta</td>
<td></td>
<td>103</td>
<td>17.0</td>
<td>1.0</td>
<td>8.3</td>
<td>0.6</td>
<td>6.6</td>
<td>0.4</td>
</tr>
<tr>
<td>Alpha + beta</td>
<td></td>
<td>118</td>
<td>17.0</td>
<td>1.0</td>
<td>8.8</td>
<td>0.5</td>
<td>7.8</td>
<td>0.5</td>
</tr>
<tr>
<td>Alpha/beta</td>
<td></td>
<td>169</td>
<td>29.0</td>
<td>1.0</td>
<td>18.5</td>
<td>0.6</td>
<td>10.0</td>
<td>0.3</td>
</tr>
</tbody>
</table>

\(^a\) As in Table 1; note all values compiled for the first \(L/2\) predictions (Methods).
• Connecting segment are very informative for contact formation,
• Evolutionary profile are crucial,
• Contact density dependent on type of protein,
• Similar accuracy but better performance for short proteins,
• α worst and α/β best,
• Of 50 % within 2 residues of observed contact,
• Correct for core, hydrophobic and regular secondary structure.
• **Limitations** : very large training set required (400,000 contacts).
Overview

- **Profcon**: Neural Network Approach
- **BetaPro**: Hierarchical Using Neural Networks, Alignments, and Graph Algorithms
- **SMURF**: Markov Random Field
- **tFolder**: Abstract template & ensemble prediction
Three-Stage Prediction of Beta-Sheets

- **Stage 1**
  Predict beta-residue pairing probabilities using 2D-Recursive Neural Networks (2D-RNN, Baldi and Pollastri, 2003)

- **Stage 2**
  Use beta-residue pairing probabilities to align beta-strands

- **Stage 3**
  Predict beta-strand pairs and beta-sheet architecture using graph algorithms
Dataset and Statistics

- Extract proteins with high resolution from Protein Data Bank (Berman et al., 2000)
- Use DSSP (Kabsch and Sander, 1983) to assign intra-chain beta-sheet structure
- Use UniqueProt (Mika and Rost, 2003) to reduce redundancy
- Use PSI-BLAST (Altschul et al., 1997) to generate profiles

### Statistics

<table>
<thead>
<tr>
<th></th>
<th>Num</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Chains</td>
<td>916</td>
</tr>
<tr>
<td>Beta Residues</td>
<td>48,996</td>
</tr>
<tr>
<td>Beta Residue Pairs</td>
<td>31,638</td>
</tr>
<tr>
<td>Beta Strands</td>
<td>10,745</td>
</tr>
<tr>
<td>Beta Strand Pairs</td>
<td>8,172</td>
</tr>
<tr>
<td>Beta Sheets</td>
<td>2,533</td>
</tr>
</tbody>
</table>
Stage 1: Prediction of Beta-Residue Pairings Using 2D-RNN

Input Matrix I (m×m)

Target / Output Matrix (m×m)

I_{ij}:

- \( X_i \) or \( X_j \) is the position of beta-residue i or j in the sequence
An Example (Target)

Protein 1VJG

Beta-Residue Pairing Map (Target Matrix)
An Example (Target)

Protein 1VJG

Beta-Residue Pairing Map (Target Matrix)
An Example (Prediction)
Stage 2: Beta-Strand Alignment

- Use output probability matrix as scoring matrix
- Dynamic programming
- Disallow gaps and use the simplified search algorithm

Total number of alignments = 2(m+n-1)
Strand Alignment and Pairing Matrix

- The alignment score is the sum of the pairing probabilities of the aligned residues
- The best alignment is the alignment with the maximum score
- Strand Pairing Matrix

Strand Pairing Matrix of 1VJG
Stage 3: Prediction of Beta-Strand Pairings and Beta-Sheet Architecture (Constraints)

(a) Seven strands of protein 1VJG in sequence order

(b) Beta-sheet topology of protein 1VJG
Stage 3: Prediction of Beta-Strand Pairings and Beta-Sheet Architecture (Constraints)

(a) Seven strands of protein 1VJG in sequence order

(b) Beta-sheet topology of protein 1VJG
Minimum Spanning Tree Like Algorithm

Strand Pairing Graph (SPG)

Strand Pairing Matrix

(a) Complete SPG
Minimum Spanning Tree Like Algorithm

Goal: Find a set of connected subgraphs that maximize the sum of the alignment scores and satisfy the constraints

Algorithm: Minimum Spanning Tree Like Algorithm

Strand Pairing Graph (SPG)

(a) Complete SPG

(b) True Weighted SPG

Strand Pairing Matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.3</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.94</td>
<td>.37</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.02</td>
<td>.02</td>
<td>.04</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.02</td>
<td>.02</td>
<td>.03</td>
<td>1.9</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.10</td>
<td>.05</td>
<td>.74</td>
<td>.04</td>
<td>.04</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.02</td>
<td>.02</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
<td>.20</td>
<td>0</td>
</tr>
</tbody>
</table>
### An Example of MST Like Algorithm

#### Strand Pairing Matrix of 1VJG

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.30</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.94</td>
<td>0.37</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.02</td>
<td>0.02</td>
<td>0.04</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>1.90</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.10</td>
<td>0.05</td>
<td>0.74</td>
<td>0.04</td>
<td>0.04</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td>0.20</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Step 1: Pair strand 4 and 5
## An Example of MST Like Algorithm

**Strand Pairing Matrix of 1VJG**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.3</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.94</td>
<td>.37</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.02</td>
<td>.02</td>
<td>.04</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.02</td>
<td>.02</td>
<td>.03</td>
<td>1.9</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.10</td>
<td>.05</td>
<td>.74</td>
<td>.04</td>
<td>.04</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.02</td>
<td>.02</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
<td>.20</td>
<td>0</td>
</tr>
</tbody>
</table>

**Step 2: Pair strand 1 and 2**

- Pair strand 1 and 2.

- Strand 1: 2
- Strand 2: 1
- N: 4, 5
An Example of MST Like Algorithm

Strand Pairing Matrix of 1VJG

Step 3: Pair strand 1 and 3
### An Example of MST Like Algorithm

#### Strand Pairing Matrix of 1VJG

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.3</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.94</td>
<td>.37</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.02</td>
<td>.02</td>
<td>.04</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.02</td>
<td>.02</td>
<td>.03</td>
<td>1.9</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.10</td>
<td>.05</td>
<td>.74</td>
<td>.04</td>
<td>.04</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.02</td>
<td>.02</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
<td>.20</td>
<td>0</td>
</tr>
</tbody>
</table>

Step 4: Pair strand 3 and 6
An Example of MST Like Algorithm

Strand Pairing Matrix of 1VJG

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.3</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.94</td>
<td>.37</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.02</td>
<td>.02</td>
<td>.04</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.02</td>
<td>.02</td>
<td>.03</td>
<td>1.9</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.10</td>
<td>.05</td>
<td>.74</td>
<td>.04</td>
<td>.04</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.02</td>
<td>.02</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
<td>.20</td>
<td>0</td>
</tr>
</tbody>
</table>

Step 5: Pair strand 6 and 7
A New Fold Example (Last CASP)

1S12 (94 residues)

True secondary structure

```
CEEEEEECCCEEEECCCCCHHHHHHHHHHHHHHHCCCCCEEEEEECCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCEEEEEECCCCCC
```

Predicted secondary structure by SSpro (Pollastri, et al., 2002)

```
CEEEEEECCCEEEECCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHEHCCCCCEEEEEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCEEEEEECCC
```

Beta Sheet Topology

Strand Pairing Matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1.71</td>
<td>.05</td>
<td>.29</td>
<td>.33</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.06</td>
<td>.41</td>
<td>.12</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>.22</td>
<td>.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>.53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

True: 1-2, 2-4, 3-4, **1-5**

Predicted: 1-2, 2-4, 3-4, **4-5**
Beta-Residue Pairing Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Specificity/Sensitivity</th>
<th>Ratio of Improvement</th>
<th>ROC Area</th>
<th>TPR at 5% FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BetaPairing</td>
<td>41%</td>
<td>17.8</td>
<td>0.86</td>
<td>58%</td>
</tr>
<tr>
<td>CMAPpro</td>
<td>27%</td>
<td>11.7</td>
<td>0.80</td>
<td>42%</td>
</tr>
</tbody>
</table>

(Pollastri and Baldi, 2002)

The accuracy of random algorithm is 2.3%.
Strand Pairing Results

- Naïve algorithm of pairing all adjacent strands
  - Specificity = 42%
  - Sensitivity = 50%
  - All strand pairs are local strand pairs.

- MST like algorithm
  - Specificity = 53%
  - Sensitivity = 59%
  - >20% correctly predicted strand pairs are non-local strand pairs.
Strand Alignment Results

On the correctly predicted strand pairs

<table>
<thead>
<tr>
<th></th>
<th>Paring Direction</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>93%</td>
<td>72%</td>
</tr>
</tbody>
</table>

On all native strand pairs

<table>
<thead>
<tr>
<th></th>
<th>Paring Direction</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>84%</td>
<td>66%</td>
</tr>
</tbody>
</table>

• The accuracy of pairing direction is 15% higher than that of the base-line algorithm.
• The alignment accuracy is significantly higher than previous methods.
Overview

- **Profcon**: Neural Network Approach
- **BetaPro**: Hierarchical Using Neural Networks, Alignments, and Graph Algorithms
- **SMURF**: Markov Random Field
- **tFolder**: Abstract template & ensemble prediction
Homology recognition of $\beta$-propeller

Seven-bladed $\beta$-propeller.
Modeling β-sheets with Markov Random Fields

HMM:

MRF:

(MRFs can model Long range dependencies of beta-strands)
## Results

<table>
<thead>
<tr>
<th>TN</th>
<th>HMMER</th>
<th>SMURF(P)</th>
<th>SMURF</th>
<th>HMMER</th>
<th>SMURF(P)</th>
<th>SMURF</th>
<th>HMMER</th>
<th>SMURF(P)</th>
<th>SMURF</th>
</tr>
</thead>
<tbody>
<tr>
<td>97%</td>
<td>52</td>
<td>20</td>
<td>80</td>
<td>80</td>
<td>23</td>
<td>87</td>
<td>0</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>96%</td>
<td>56</td>
<td>24</td>
<td>80</td>
<td>80</td>
<td>33</td>
<td>87</td>
<td>20</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>95%</td>
<td>64</td>
<td>36</td>
<td>80</td>
<td>87</td>
<td>47</td>
<td>93</td>
<td>20</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>94%</td>
<td>68</td>
<td>36</td>
<td>84</td>
<td>90</td>
<td>53</td>
<td>93</td>
<td>40</td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td>93%</td>
<td>68</td>
<td>48</td>
<td>84</td>
<td>90</td>
<td>53</td>
<td>97</td>
<td>40</td>
<td>100</td>
<td>40</td>
</tr>
<tr>
<td>92%</td>
<td>68</td>
<td>60</td>
<td>88</td>
<td>90</td>
<td>57</td>
<td>97</td>
<td>40</td>
<td>100</td>
<td>40</td>
</tr>
<tr>
<td>91%</td>
<td>68</td>
<td>60</td>
<td>92</td>
<td>90</td>
<td>57</td>
<td>97</td>
<td>40</td>
<td>100</td>
<td>40</td>
</tr>
<tr>
<td>90%</td>
<td>68</td>
<td>60</td>
<td>92</td>
<td>93</td>
<td>57</td>
<td>100</td>
<td>60</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Overview

- **Profcon**: Neural Network Approach
- **BetaPro**: Hierarchical Using Neural Networks, Alignments, and Graph Algorithms
- **SMURF**: Markov Random Field
- **tFolder**: Abstract template & ensemble prediction
Classical View of Structural Biology

ATSTVTGGYAQSDAQGQMNK
MGGFNLKYRYEEDNSPLGVIGSF
TYTEKSRTASSGDYNKNQYYGITA
GPAAYRINDWASIYGVGVGYG
KFQTEYPTYKNDSGYFSGYGA
GLQFNPMENVALDFSYEQSRI
VDVGWTIAIGVGYRF
Ensemble modeling

- Single structure prediction - Classical view.

- Ensemble prediction - Modern view.

**Our approach:** Compute a realistic ensemble representation of the structure landscape.
Ensemble approach enables:

1. Protein structure prediction.  
   “Ensembles provide realistic structure prediction”  
   (Waldispühl et al., Proteins, 2008; O’Donnell et al., ISMB 2011)

2. Folding pathway prediction.  
   “Ensembles enable fast prediction of folding dynamics”  
   (Shenker et al., RECOMB, 2011)
Modeling ensembles of structures

Gibbs Measure:

\[ P(S) \propto e^{\frac{-E(S)}{RT}} \]

Structure probability is proportional to the exponential of its energy.

This defines a *Boltzmann distribution* enabling:
- to compute statistics.
- to sample Structures.
Ensemble approach enables:

1. **Protein structure prediction.**
   "Ensembles provide realistic structure prediction"  
   (Waldispühl *et al.*, Proteins, 2008; O’Donnell *et al.*, ISMB 2011)

2. Folding pathway prediction.
   "Ensembles enable fast prediction of folding dynamics"  
   (Shenker *et al.*, RECOMB, 2011)
Transmembrane $\beta$-barrel proteins

- Found in outer-membranes.
- Wide variety of functions.

- Difficult to solve with X-Ray/NMR techniques,
  - Only few non-homologous structures in PDB.

Ompx, E. Coli (Vogt & Schulz, 1999)

FhuA, E. coli (Pawelek et al., Science, 2006)
1. The barrel is decomposed in a sum of β–strand pairs.

2. Inclination modeled using strand extensions.

(Waldispuhl et al., Proteins, 2006)
Exploring the TMB folding landscape

1. Initialization: create strand pairs.

2. Chaining: concatenate strand pairs to build $\beta$-sheets.

3. Closure: Pair first and last strand pair of the $\beta$-sheet.

(Waldispühl et al., *Proteins*, 2006)
Energy model

- Classical approach: Residue contacts.

Energy of the structure is the sum of the energies of all contacts.

- Our new Concept: **Stacking pairs.**

\[ E(i, j, x | i + 2, j + 2) = -RT \log(p_{i,j,x|i+2,j+2}) - RT \log(Q_{mb}) \]

- Computed from globular proteins
- Distinguish environment

(Waldispuhl et al., *Proteins*, 2008)
Stochastic Contact Maps

Contact probability:

\[ p(i,j) \propto \sum_{(i,j) \in S} e^{-\frac{E(S)}{RT}} \]

Backtrack dynamic tables \( O(n^3) \).

Upper triangle: Membrane
Lower triangle: Channel

(Waldispuhl et al., Proteins, 2008)
Stochastic Contact Maps

Red: Crystal structure
Green: Single structure prediction

Initial α-helix stabilizes PagP (Huysmans et al., 2007)

Comprehensive representation of ensemble of folds

(Waldispühl et al., Proteins, 2008)
**Contact Profiles**

**B-factor**: experimental measure of the flexibility of residues.

Per residue contact probability correlates with B-factor

\[
P_c(i) = 1 - \sum_{j} p(i,j)
\]

**Direct Prediction of experimental measures.**

(Waldispühl et al., *Proteins*, 2008)
Modeling $\beta$-sheets

$\beta$-sheet topologies modeled as signed permutations

Orientation:

$+$

$-$
Modeling $\beta$-sheets

$\beta$-sheet model explicitly incorporates inter-strand residue interaction with side-chain orientation.

Orientation:

+  

Residue with side-chain orientation

Inter-strand residue contact
Recursive enumeration of $\beta$-sheets

Expand dynamic tables to allow strand insertion:

Signed permutations define the order of strand insertions.
Contact prediction benchmark

Benchmark 12 non-homologous protein data set

<table>
<thead>
<tr>
<th>Separation</th>
<th>Method</th>
<th>Accuracy</th>
<th>Coverage</th>
<th>Method</th>
<th>Accuracy</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;= 12</td>
<td>tFolder</td>
<td>0.27</td>
<td>0.27</td>
<td>tFolder</td>
<td>0.23</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>BETApro</td>
<td>0.22</td>
<td>0.40</td>
<td>BETApro</td>
<td>0.05</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>SVMcon</td>
<td>0.32</td>
<td>0.31</td>
<td>SVMcon</td>
<td>0.24</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Accuracy = correct/total predicted   Coverage = correct/total native

Performance not affected by residue sequence separation
EVcoupling

(Marks et al., 2011)