Protein tertiary structure prediction with new machine learning approaches

Rui Kuang

Department of Computer Science
Columbia University

Supervisor: Jason Weston(NEC) and Christina Leslie(Columbia)

NEC summer internship talk, August 30th, 2005
Agenda

1. Introduction to protein structure
2. Protein backbone angle prediction with structured output learning
3. Protein domain detection based on protein structural classification
4. Discussion
Part 1: Protein structure

- **Protein** – Derived from Greek word *proteios* meaning "of the first rank" in 1838 by Jöns J. Berzelius
- Crucial in all biological processes
- Function depends on structure (structure can help us to understand function)
- Determination of protein structures is time consuming and expensive
How to describe protein structure

- Primary structure: amino acid sequence
- Secondary structure: local structure elements
- Tertiary structure: packing and arrangement of secondary structure, also called domain
- Quaternary structure: arrangement of several polypeptide chains
Describe protein tertiary structure by protein backbone angles

**Phi-Psi Angles**

\[
(\Phi_1, \Psi_1) \\
(\Phi_2, \Psi_2) \\
(\Phi_3, \Psi_3) \\
(\Phi_4, \Psi_4) \\
(\Phi_5, \Psi_5) \\
(\Phi_6, \Psi_6) \\
(\Phi_7, \Psi_7) \\
(\Phi_8, \Psi_8) \\
......
\]

3-D structure

*(Too complicated to predict!)*

Simplify
Discretization of Phi-Psi angles: conformational states

Oliver et al. (Journal of Molecular Biology, 1997)
Protein blocks

16 small prototypes (a-p) of local protein structures of 5 residue length, clustered from Phi-Psi angles

De Brevern et al. (Protein Science, 2002)
Summary: representations of 3-D protein structure

**Phi-Psi Angles**

- $(\Phi_1, \Psi_1)$
- $(\Phi_2, \Psi_2)$
- $(\Phi_3, \Psi_3)$
- $(\Phi_4, \Psi_4)$
- $(\Phi_5, \Psi_5)$
- $(\Phi_6, \Psi_6)$
- $(\Phi_7, \Psi_7)$
- $(\Phi_8, \Psi_8)$

**Conformational States:**

AAAGBBBBBBBGEBBBBB...

**Protein Blocks:**

ammmalpppmmmlmlbb...
Protein domains

• A polypeptide chain or a part of a polypeptide chain that can fold independently into a stable tertiary structure.
Part 2:
Prediction of protein backbone angle with structured output learning
Naïve window-based approach

Encode each position independently with sequence information within a length-k window.

Conformational States

| A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| -3 | -4 | -4 | -4 | -3 | -4 | -4 | -2 | -1 | -1 | -4 | -1 | 8 | -5 | -3 | -3 | 0 | 2 | -2 |
| 0 | -1 | -1 | -4 | 3 | 4 | 1 | -1 | -4 | -4 | 0 | -3 | -4 | -2 | -1 | -2 | -4 | -3 | -3 |
| 0 | -1 | 2 | 1 | -3 | 4 | 0 | -1 | -2 | -4 | -3 | 1 | -2 | -4 | -2 | 2 | 0 | -4 | -3 | -3 |
| -2 | -3 | -4 | -5 | -2 | -3 | -4 | -6 | -4 | 0 | 6 | 0 | 0 | -1 | -4 | -3 | -2 | -4 | -2 | 0 |
| 0 | -3 | -1 | -2 | -3 | 0 | -2 | 4 | -3 | -3 | 0 | -2 | -2 | -4 | -3 | 3 | 1 | -4 | -4 | -3 |
| 0 | 2 | 0 | 4 | -4 | 1 | 2 | 1 | -2 | -4 | -4 | 0 | -3 | -4 | -3 | 1 | -2 | -5 | -4 | -4 |
| -1 | 5 | 3 | -2 | -4 | -1 | 1 | -2 | -1 | -4 | 1 | -3 | -4 | -3 | 1 | -2 | -5 | -4 | -4 |

To SVM

Predictions are independent.

We are neighbors! We have dependency!

Kuang, Leslie and Yang et al. (Bioinformatics, 2004)
2-Stage window-based approach

- Take the prediction of the naïve window-based approach as input to a second sets of SVMs.
- Ideally this smoothing step can correct some wrong predictions.
Topographic SVM

- Training with profiles + true labels
- Iteratively update the predictions in the testing phase.

Mohr and Obermayer et al. (NIPS, 2004)
Struct-SVM

• Training: make joint feature mapping $\Psi(x, y)$ and apply large margin principle for the difference between the feature mapping of correct label and of wrong label.

\[ \forall i \in \{1, \ldots, n\} : \max_{y \in \mathcal{Y} \setminus y_i} \{\langle w, \Psi(x_i, y) \rangle\} \leq \langle w, \Psi(x_i, y_i) \rangle. \]

• This is equivalent to the following optimization problem

\[ \min_w \frac{1}{2} \|w\|^2 \quad \text{s.t.} \quad \forall i, \forall y \in \mathcal{Y} \setminus y_i : \langle w, \Psi(x_i, y_i) - \Psi(x_i, y) \rangle \geq 1. \]

• Testing: a pre-image problem

\[ f(x; w) = \arg\max_{y \in \mathcal{Y}} \langle w, \Psi(x, y) \rangle. \]

Tsochantaridis (ICML, 2004)
Pre-image for Labeling Sequences

- Hidden-Markov kernel

\[
F(x, y) = F_1(x, y) + F_2(x, y)
\]

\[
F_1(x, y) = \sum_{\sigma, \tau} \sum_{i, \bar{y}} \alpha_i(\bar{y}) \sum_t [[y^{t-1} = \sigma \land \bar{y}^t = \tau]] \sum_s [[y^{s-1} = \sigma \land y^s = \tau]]
\]

\[
F_2(x, y) = \sum_{s, \sigma} [[y^s = \sigma]] \sum_{i, t} \sum_y [[y^t = \sigma]] \alpha_i(y) k(x^s, x_i^t)
\]

- Pre-image is equivalent to Viterbi-decoding of a HMM built from support vectors

Altun et al. (ICML 2003)
Preliminary Results

- **Prediction of Conformational States:** 697 sequences of 97,365 amino acids with sequence identity < 25%
- **Prediction of Protein Blocks:** 675 sequences of 146,978 amino acids with sequence identity < 30%

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy (Conformational States)</th>
<th>Accuracy (Protein Blocks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>State of art</td>
<td>75.0%</td>
<td>40.3%</td>
</tr>
<tr>
<td>Naïve window-based approach</td>
<td>75.0%</td>
<td>57.7%</td>
</tr>
<tr>
<td>2-Stage window-based approach</td>
<td>76.0%&gt;</td>
<td>59.5%&gt;</td>
</tr>
<tr>
<td>Topographic SVM</td>
<td>75.3%</td>
<td>58.4%</td>
</tr>
<tr>
<td>SVM for structured output</td>
<td>70.0%&gt;</td>
<td>50%&gt;</td>
</tr>
</tbody>
</table>
Part 3:
Protein domain detection based on protein structural classification
Protein structural classification

**Family**: Sequence identity > 30% or functions and structures are very similar

**Superfamily**: low sequence similarity but functional features suggest probable common evolutionary origin

**Common fold**: same major secondary structures in the same arrangement with the same topological connections

Murzin et al. (Journal of Molecular Biology, 1995)
Spectrum kernel

- Feature map indexed by all possible k-length subsequences ("k-mers") from alphabet $\Sigma$ of amino acids, $|\Sigma| = 20$

$$K(Q_1, Q_2) = \langle \ldots 1 \ldots 0 \ldots 1 \ldots 0 \ldots 1 \ldots 0 \ldots 1 \ldots 2 \rangle, \langle \ldots 1 \ldots 1 \ldots 1 \ldots 0 \ldots 1 \ldots 0 \ldots 1 \ldots 1 \ldots 0 \rangle \geq 3$$

Leslie et al. (PSB, 2002)
Profile kernel

- Use profile \( P(x) = \{ p_j(b), b \in \Sigma, j = 1 \ldots |x| \} \) to define position-dependent mutation neighborhoods:
- E.g. \( k=3, \sigma=5 \) and a profile of negative log probabilities

\[
M_{(k,\sigma)}(P(x[j+1:j+k])) = \{ b_1 b_2 \ldots b_k : -\sum_i \log(p_{j+i}(b_i)) < \sigma \}
\]

<table>
<thead>
<tr>
<th>A</th>
<th>K</th>
<th>Q</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>K</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Q</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Y</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

\( (2+1+1<\sigma) \quad (1+1+1<\sigma) \quad (1+1+2<\sigma) \)

Kuang and Leslie et al. (JBCB, 2005)
Positional classification scores

A simple probabilistic model to detect domains:

\[
P(S, E \mid F) = \overline{P}(s_0, s_1 - 1 \mid F) \cdot P(s_1, e_1 \mid F) \cdot \overline{P}(e_1 + 1, s_2 - 1 \mid F) \cdot \\
P(s_2, e_2 \mid F) \cdot \overline{P}(e_2 + 1, s_3 - 1 \mid F) \cdots P(s_n, e_n \mid F) \cdot \overline{P}(e_n + 1, \mid F \parallel F)
\]
Experiments

1. Dataset
   - 7,329 sequences from SCOP 1.59.
   - Sequence identity less than 95%.

2. Preliminary Results (with a simplified model)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Accuracy</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain positions</td>
<td>73.2%</td>
<td>73.1%</td>
</tr>
<tr>
<td>Domain start</td>
<td>51.1%</td>
<td>36.0%</td>
</tr>
<tr>
<td>Domain end</td>
<td>31.1%</td>
<td>21.9%</td>
</tr>
</tbody>
</table>
Part 4: Discussion

• Dependency between conformational states or protein blocks does not help much in the 2-stage window-based approach.

• Struct-SVM does not scale very well for large problems. Perceptron training may speed up the training stage.

• A proper probabilistic model is needed for detecting domain boundaries from positional classification scores
Acknowledgement

• William Stafford Noble
  Genome Science Department, University of Washington

• Asa Ben-Hur
  Genome Science Department, University of Washington

• An-Suei Yang
  Genome Research Center, Academia Sinica of Taiwan

• Yasemin Altun
  Toyota Technological Institute at Chicago

• Thorsten Joachims
  Computer Science Department, Cornell University