A Graphical Model for Protein Secondary Structure Prediction

Wei Chu, Zoubin Ghahramani, David L. Wild
Gatsby Computational Neuroscience Unit, University College London
Keck Graduate Institute of Applied Life Sciences, USA
chuwei@gatsby.ucl.ac.uk
www.gatsby.ucl.ac.uk/~chuwei/

2004/07/07

Gatsby Computational Neuroscience Unit
Protein Structures

Primary Structure $\rightarrow$ Secondary Structure $\rightarrow$ Tertiary Structure
Protein Secondary Structure Prediction

- Discriminant approach with neural networks,
  - seminal work by Qian and Sejnowski (1988);
  - PHD (Rost and Sander, 1993) employed evolutionary information in the form of profiles derived from multiple sequence alignment;
  - another type of alignment profile, position-specific scoring matrices (PSSM) derived by PSI-BLAST (Altschul et al. 1997), has been used in neural network methods (Jones, 1999; Cuff and Barton, 2000).

- Generative modelling,
  - Delcher et al. (1993) applied Hidden Markov Models (HMM)
  - Schmidler (2002) presented a segmental semi-Markov model (SSMM) using sequence information only. The prediction accuracy of the SSMM still falls short of the best contemporary discriminative methods.
Our Approach

• present a probabilistic generative graphical model that extends segmental semi-Markov models (SSMM) to exploit multiple sequence alignment profiles which contain information from evolutionarily related sequences;

• propose a novel parameterized model as the likelihood function to capture the segmental conformation;

• incorporate the information from long range interactions in $\beta$-sheets for contact map prediction;

• predict structure of novel protein using Bayesian inference, e.g. belief propagation and Markov chain Monte Carlo methods.
Multiple Sequence Alignment Profiles

- The alignment profile $O = [O_1, O_2, \ldots, O_i, \ldots, O_n]$ is a sequence of $20 \times 1$ vectors, where $O_i$ contains the occurrence counts for the 20 amino acids at location $i$ which can be regarded as a realization of a multinomial random variable;
- A set of segmental variables, $(m, e, T)$, where $m$ is the number of segments, the segmental endpoints $e = [e_1, e_2, \ldots, e_m]$ and the segment types $T = [T_1, T_2, \ldots, T_m]$. 
Segmental Semi-Markov Models
Bayesian Framework

- Prior probability $\mathcal{P}(m, e, T)$:

$$
\mathcal{P}(m, e, T) = \mathcal{P}(m) \prod_{i=1}^{m} \mathcal{P}(e_i|e_{i-1}, T_i) \mathcal{P}(T_i|T_{i-1})
$$

- a uniform prior for the number of segments $m$, $\mathcal{P}(m) \propto 1$
- segmental length distribution $\mathcal{P}(e_i|e_{i-1}, T_i) = \mathcal{P}(l_i = e_i - e_{i-1}|T_i)$
- segment type transition probabilities $\mathcal{P}(T_i|T_{i-1})$. 
A Graph for $\mathcal{P}(l_i = e_i - e_{i-1}|T_i)$

The Distributions of Segmental Length

![Graph showing the distributions of segmental lengths for different states: Helix, Strand, and Coil. The x-axis represents the segmental length, and the y-axis represents the probability.]
A Graph for $\mathcal{P}(T_i|T_{i-1})$

Segment Type Transition Probabilities

- Strand to Helix: 0.0541
- Strand to Coil: 0.0200
- Helix to Strand: 0.0060
- Helix to Helix: 0.0371
- Helix to Coil: 0.0013
- Coil to Strand: 0.5269
- Coil to Helix: 0.9399
- Coil to Coil: 0.9429
Likelihood Evaluation

- Likelihood $\mathcal{P}(O|m, e, T)$:

\[
\mathcal{P}(O|m, e, T) = \prod_{i=1}^{m} \prod_{k=e_{i-1}+1}^{e_{i}} \mathcal{P}(O_k|O_{[1:k-1]}, T_i)
\]

$m$ - the number of segments;
$e$ - the segmental endpoints;
$T$ - the segment types;
$O$ - the observations;
$O_k$ - the observation at the $k$-th residue, i.e. a multinomial realization.
Dependency Window

$T_1 = C$  $T_2 = E$  $T_m = C$

$I_1 = 3$  $I_2 = 4$  $I_m = 4$

$O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8, O_9, O_{10}, O_{11}$

$θ_{n-6}, θ_{n-5}, θ_{n-4}, θ_{n-3}, θ_{n-2}, θ_{n-1}$

$W^2_{inter}, W^3_{inter}, W^1_{intra}$

$O_{n-6}, O_{n-5}, O_{n-4}, O_{n-3}, O_{n-2}, O_{n-1}$

$O_n$
Individual Likelihood

This is a **Dirichlet-Multinomial** distribution.

\[
\mathcal{P}(O_k|O_{[1:k-1]}, T_i) = \int_{\theta_k} \mathcal{P}(O_k|\theta_k, T_i) \mathcal{P}(\theta_k|O_{[1:k-1]}, T_i) \, d\theta_k
\]

- **Multinomial**: \( \mathcal{P}(O_k|\theta_k, T_i) = \frac{(\sum_a O^a_k)!}{\prod_a O^a_k!} \prod_{a \in A} (\theta^a_k)^{O^a_k} \)

- **Dirichlet Prior**: \( \mathcal{P}(\theta_k|O_{[1:k-1]}, T_i) = \frac{\Gamma(\sum_a \gamma^a_k)}{\prod_a \Gamma(\gamma^a_k)} \prod_{a \in A} (\theta^a_k)^{\gamma^a_k - 1} \)

- **Weights**: \( \gamma_k = W_{\text{cap}} + \sum_{j=1}^{\ell_k} W_{\text{intra}}^j \cdot O_{k-j} + \sum_{j=\ell_k+1}^{\ell} W_{\text{inter}}^j \cdot O_{k-j} \).
Posterior Probability

The posterior probability can be written as

\[
P(m, e, T|O) \propto P(O|m, e, T) \cdot P(m, e, T)
\]

- **MAP**: the most likely segmentations, \( \arg \max_{(m, e, T)} P(m, e, T|O) \); (Viterbi algorithm)

- **Marginal**: the marginal posterior distribution of the segment type at each residue, \( P(T_{O_i}|O) \), where \( T_{O_i} \) denotes the segment type at the \( i \)-th residue. (Forward-Backward Algorithm)
Parameter Estimate

- The parameters for discrete distributions, including the state transition probability $P(T_i|T_{i-1})$ and segmental length distribution $P(e_i|e_{i-1}, T_i)$ can be estimated by their relative frequency of occurrence in the training dataset.

- The weights in segmental likelihood, which consist of three subsets for different segmental types, i.e. $\{W_\tau\}$ with $\tau \in \{H, E, C\}$.

$$\arg \max_{W_\tau} P(\{O, m, e, T\}|W_\tau)P(W_\tau)$$

where $P(W_\tau)$ is the prior probability usually specified by $P(W_\tau) \propto \exp(-\frac{C_\tau}{2}\|W_\tau\|_2^2)$ with $C_\tau \geq 0$. 
### 7-fold Cross Validation on 480 chains of CB513

<table>
<thead>
<tr>
<th></th>
<th>Sequence Only</th>
<th>with MSAP</th>
<th>with PSSM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAP</td>
<td>MARG</td>
<td>MAP</td>
</tr>
<tr>
<td>$Q_3$</td>
<td>59.23%</td>
<td>65.08%</td>
<td>68.34%</td>
</tr>
<tr>
<td>$Q_{obs}^H$</td>
<td>66.34%</td>
<td>66.73%</td>
<td>78.28%</td>
</tr>
<tr>
<td>$Q_{obs}^E$</td>
<td>20.74%</td>
<td>46.32%</td>
<td>42.55%</td>
</tr>
<tr>
<td>$Q_{obs}^C$</td>
<td>72.80%</td>
<td>73.19%</td>
<td>73.14%</td>
</tr>
<tr>
<td>$Q_{pred}^H$</td>
<td>61.87%</td>
<td>68.64%</td>
<td>70.56%</td>
</tr>
<tr>
<td>$Q_{pred}^E$</td>
<td>56.45%</td>
<td>58.88%</td>
<td>69.35%</td>
</tr>
<tr>
<td>$Q_{pred}^C$</td>
<td>57.77%</td>
<td>64.72%</td>
<td>66.21%</td>
</tr>
</tbody>
</table>

“Sequence Only” denotes the algorithm of Schmidler et al. 2000; “MSAP” denotes our approach using multiple sequence alignment profiles; “PSSM” denotes using position specific score matrices. $Q_3$ denotes the overall accuracy. $Q_{obs} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}$, and $Q_{pred} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}$. MAP denotes the most probable posterior estimate, while MARG denotes marginal posterior mode estimate.
Predictive Results on the Protein Data of CASP

<table>
<thead>
<tr>
<th></th>
<th>CASP2</th>
<th>CASP3</th>
<th>CASP4</th>
<th>CASP5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(20 chains)</td>
<td>(36 chains)</td>
<td>(40 chains)</td>
<td>(56 chains)</td>
</tr>
<tr>
<td>$Q_3$</td>
<td>73.40%</td>
<td>71.12%</td>
<td>74.32%</td>
<td>74.03%</td>
</tr>
<tr>
<td>$Q_{obs}$</td>
<td>76.62%</td>
<td>73.12%</td>
<td>80.22%</td>
<td>80.43%</td>
</tr>
<tr>
<td>$Q_{obs}$</td>
<td>61.29%</td>
<td>56.35%</td>
<td>57.81%</td>
<td>59.52%</td>
</tr>
<tr>
<td>$Q_{obs}$</td>
<td>77.73%</td>
<td>78.88%</td>
<td>78.00%</td>
<td>76.81%</td>
</tr>
<tr>
<td>$Q_{pred}$</td>
<td>79.71%</td>
<td>74.91%</td>
<td>81.33%</td>
<td>76.95%</td>
</tr>
<tr>
<td>$Q_{pred}$</td>
<td>76.48%</td>
<td>78.39%</td>
<td>76.19%</td>
<td>78.10%</td>
</tr>
<tr>
<td>$Q_{pred}$</td>
<td>67.36%</td>
<td>65.99%</td>
<td>67.28%</td>
<td>69.88%</td>
</tr>
</tbody>
</table>

The predictive results of marginal posterior mode estimate (MARG) using position specific score matrices (PSSM) as alignment profile. $Q_3$ is the overall accuracy, $Q_{obs} = \frac{TruePositive}{TruePositive+FalseNegative}$, and $Q_{pred} = \frac{TruePositive}{TruePositive+FalsePositive}$. 
Long-range Interactions in $\beta$-sheets

The $\beta$-sheet space is the set of all the possible combinations of $\beta$-sheets; A set of interaction variables, $I$, to describe one possible case.
Enhanced Framework

- Prior: \( \mathcal{P}(m, e, T, I) = \mathcal{P}(I|m, e, T) \mathcal{P}(m, e, T) \), where \( \mathcal{P}(I|m, e, T) \) is the distribution over \( \beta \)-sheet space;
- Likelihood: \( \mathcal{P}(O|m, e, T, I) \);
- Posterior: \( \mathcal{P}(m, e, T, I|O) \propto \mathcal{P}(O|m, e, T, I) \mathcal{P}(m, e, T) \);
- Metropolis-Hasting scheme with reversible-jumps (Green, 1995) can be used to collect samples in \( \mathcal{P}(m, e, T|O) = \sum_I \mathcal{P}(m, e, T, I|O) \).
Contact Maps

Contact Map

Protein Structure
Inference on Contact Maps

- **$\beta$-sheet contact map** is specified by the interaction set $\mathcal{I}$ as a $n \times n$ matrix $C$ whose $ij$-th entry $C_{ij}$ is defined as

$$C_{ij}(\mathcal{I}) = \begin{cases} 1 & \text{if } O_i \text{ and } O_j \text{ are paired in the interaction set } \mathcal{I}; \\ 0 & \text{otherwise} \end{cases}$$

- Our interest is the marginal $P(C_{ij} = 1) = \sum_{m,e,T,I} C_{ij}(\mathcal{I}) P(m, e, T, \mathcal{I}|O)$

- Sampling estimate is

$$P(C_{ij} = 1) \approx \frac{1}{N} \sum_{\{m,e,T\}} \sum_{\{\mathcal{I}\}} C_{ij}(\mathcal{I}) \frac{P(O|m, e, T, \mathcal{I})}{\sum_{\{\mathcal{I}\}} P(O|m, e, T, \mathcal{I})}$$

by using samples $\{\mathcal{I}\} \sim P(\mathcal{I}|m, e, T)$ and $\{m, e, T\} \sim P(m, e, T|O)$. 
An Example

True $\beta$-sheet Contact Map

Predictive $\beta$-sheet Contact Map
Summary

- A graphical model with a novel parametric likelihood function is proposed to exploit the information in alignment profiles;

- Contact maps can be inferred in the Bayesian segmental framework by incorporating long range interaction information;

- The numerical results show the generalization performance of this graphical model is competitive with other contemporary methods;

- As a future work, the graphical model could be developed for tertiary structure prediction with the inclusion of dihedral angles.
Acknowledgement

- This work is supported by the National Institutes of Health (NIH) and its National Institute of General Medical Sciences (NIGMS) division under Grant Number 1 P01 GM63208 (NIH/NIGMS grant title: Tools and Data Resources in Support of Structural Genomics);