

A Graphical Model for Protein Secondary Structure Prediction

Wei Chu, Zoubin Ghahramani, David L. Wild

Gatsby Computational Neuroscience Unit, University College London

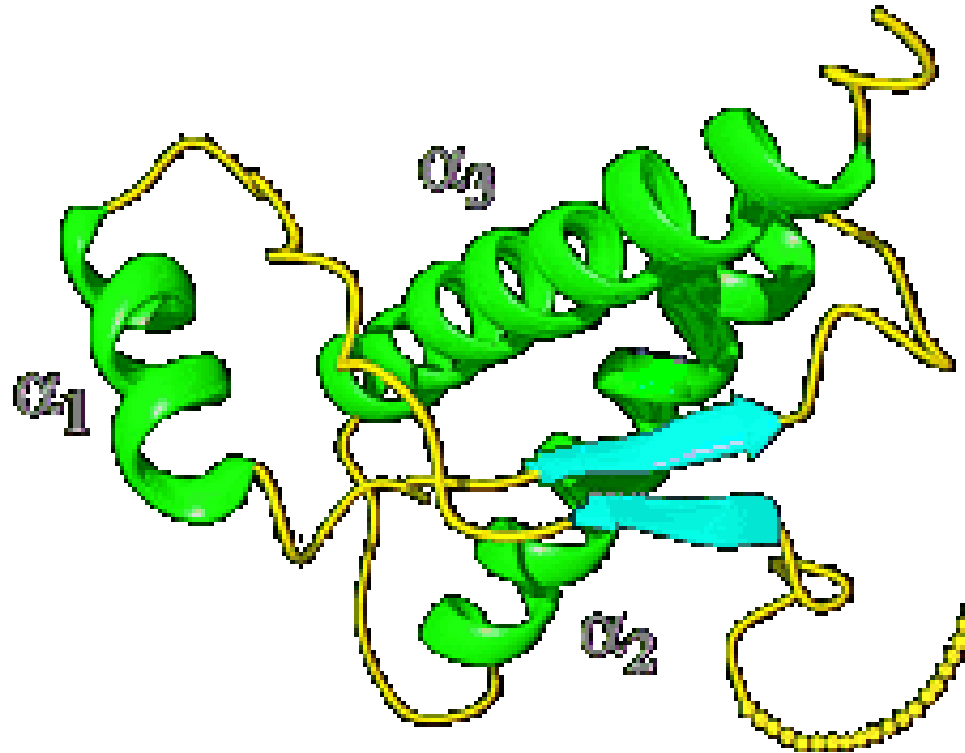
Keck Graduate Institute of Applied Life Sciences, USA

`chuwei@gatsby.ucl.ac.uk`

`www.gatsby.ucl.ac.uk/~chuwei/`

2004/07/07

Protein Structures



Primary Structure \rightarrow Secondary Structure \rightarrow Tertiary Structure

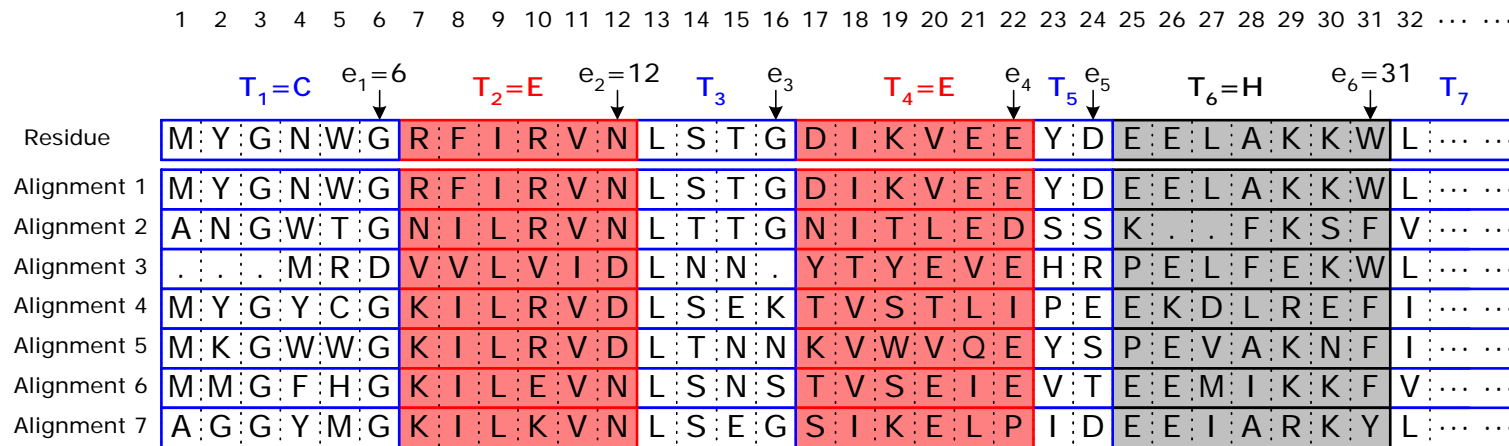
Protein Secondary Structure Prediction

- 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

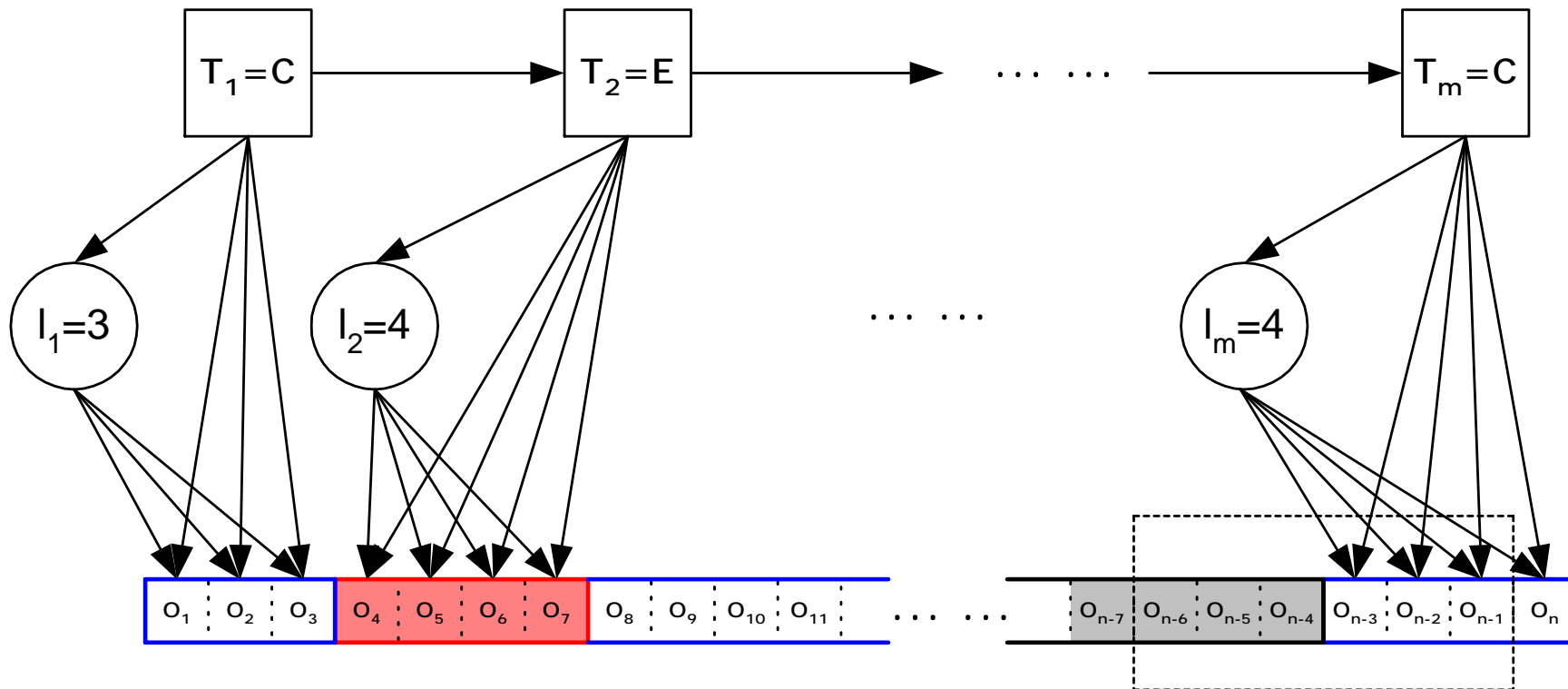
- present a probabilistic generative graphical model that extends segmental semi-Markov models (SSMM) to exploit multiple sequence alignment profiles which contain information from evolutionarily related sequences;
- propose a novel parameterized model as the likelihood function to capture the segmental conformation;
- incorporate the information from long range interactions in β -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 $O = [O_1, O_2, \dots, O_i, \dots, O_n]$ is a sequence of 20×1 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 $e = [e_1, e_2, \dots, e_m]$ and the segment types $T = [T_1, T_2, \dots, T_m]$.

Segmental Semi-Markov Models



Bayesian Framework

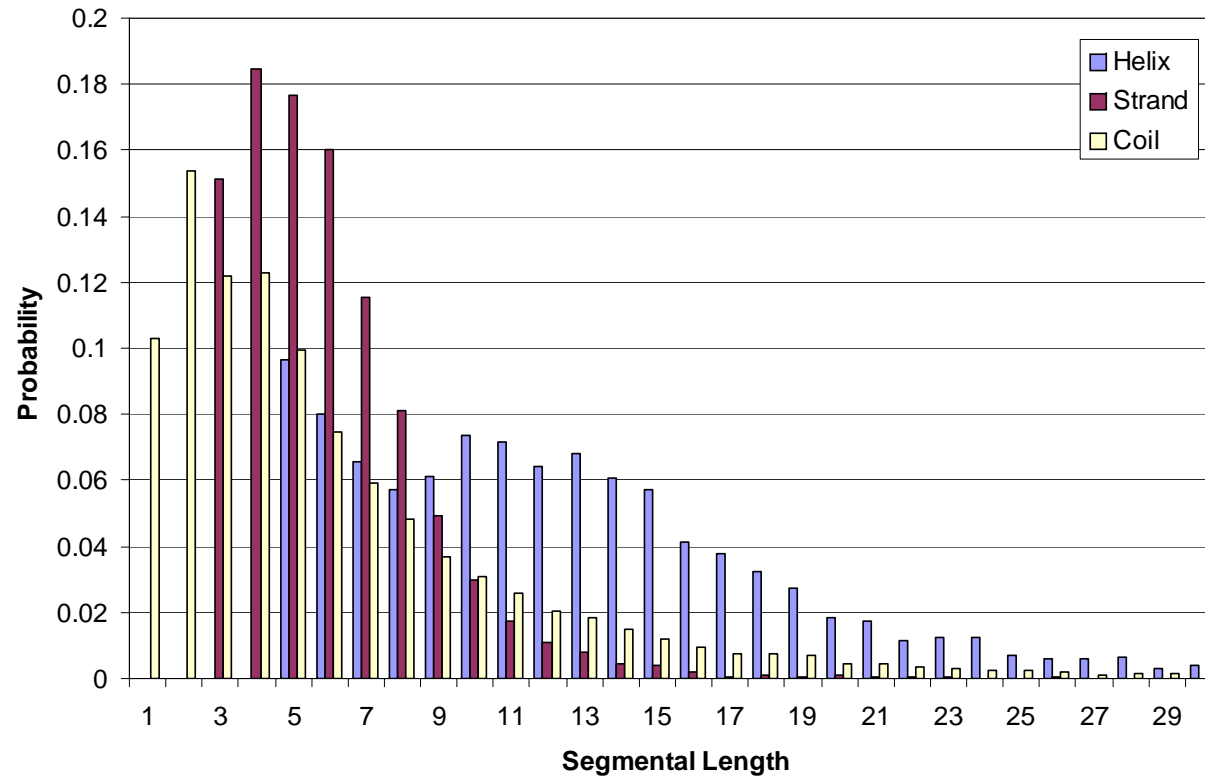
- Prior probability $\mathcal{P}(m, e, T)$:

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

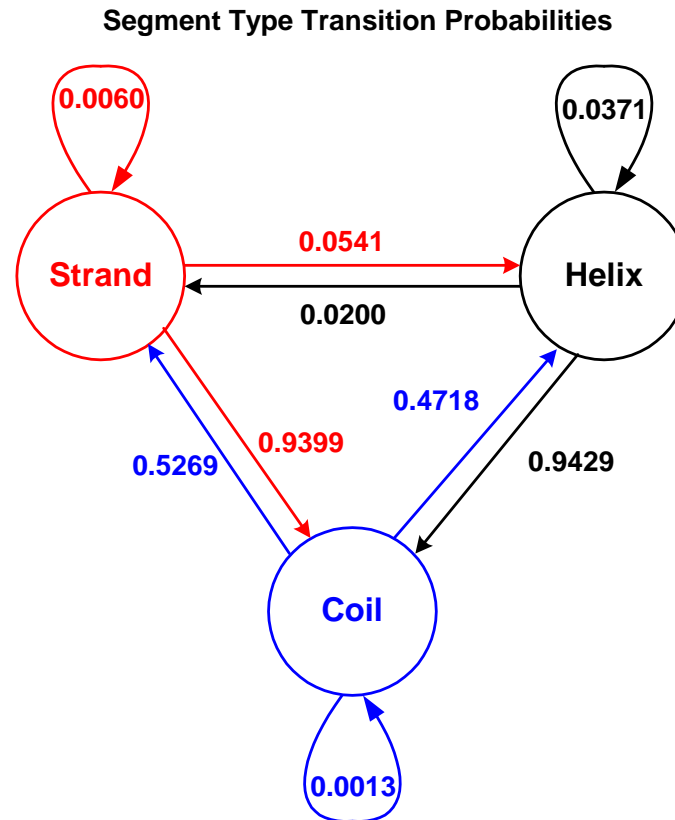
- a uniform prior for the number of segments m , $\mathcal{P}(m) \propto 1$
- segmental length distribution $\mathcal{P}(e_i | e_{i-1}, T_i) = \mathcal{P}(l_i = e_i - e_{i-1} | T_i)$;
- segment type transition probabilities $\mathcal{P}(T_i | T_{i-1})$.

A Graph for $\mathcal{P}(l_i = e_i - e_{i-1} | T_i)$

The Distributions of Segmental Length



A Graph for $\mathcal{P}(T_i|T_{i-1})$



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;

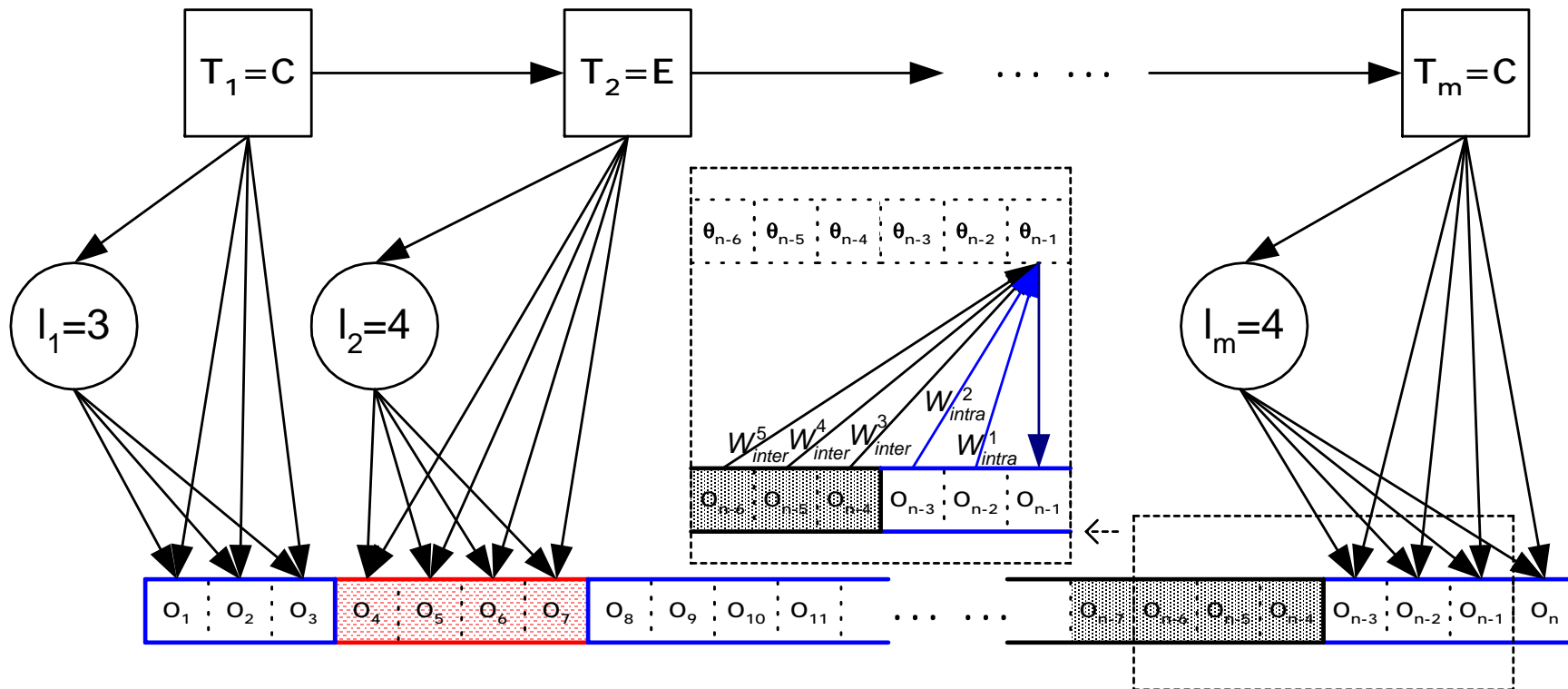
e - the segmental endpoints;

T - the segment types;

O - the observations;

O_k - the observation at the k -th residue, i.e. a multinomial realization.

Dependency Window



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: $\mathcal{P}(O_k | \theta_k, T_i) = \frac{(\sum_a O_k^a)!}{\prod_a O_k^a!} \prod_{a \in \mathcal{A}} (\theta_k^a)^{O_k^a}$
- Dirichlet Prior: $\mathcal{P}(\theta_k | O_{[1:k-1]}, T_i) = \frac{\Gamma(\sum_a \gamma_k^a)}{\prod_a \Gamma(\gamma_k^a)} \prod_{a \in \mathcal{A}} (\theta_k^a)^{\gamma_k^a - 1}$
- Weights: $\gamma_k = W_{cap} + \sum_{j=1}^{\ell_k} W_{intra}^j \cdot O_{k-j} + \sum_{j=\ell_k+1}^{\ell} W_{inter}^j \cdot O_{k-j}$.

Posterior Probability

The posterior probability can be written as

$$\underbrace{\mathcal{P}(m, e, T|O)}_{\text{Posterior}} \propto \underbrace{\mathcal{P}(O|m, e, T)}_{\text{Likelihood}} \cdot \underbrace{\mathcal{P}(m, e, T)}_{\text{Prior}}$$

- **MAP:** the most likely segmentations, $\arg \max_{(m, e, T)} \mathcal{P}(m, e, T|O)$;
(*Viterbi algorithm*)
- **Marginal:** the marginal posterior distribution of the segment type at each residue, $\mathcal{P}(T_{O_i}|O)$, 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 $\mathcal{P}(T_i|T_{i-1})$ and segmental length distribution $\mathcal{P}(e_i|e_{i-1}, T_i)$ 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. $\{\mathbf{W}_\tau\}$ with $\tau \in \{H, E, C\}$.

$$\arg \max_{\mathbf{W}_\tau} \mathcal{P}(\{O, m, e, T\} | \mathbf{W}_\tau) \mathcal{P}(\mathbf{W}_\tau)$$

where $\mathcal{P}(\mathbf{W}_\tau)$ is the prior probability usually specified by $\mathcal{P}(\mathbf{W}_\tau) \propto \exp(-\frac{C_\tau}{2} \|\mathbf{W}_\tau\|_2^2)$ with $C_\tau \geq 0$.

7-fold Cross Validation on 480 chains of CB513

	Sequence Only		with MSAP		with PSSM	
	MAP	MARG	MAP	MARG	MAP	MARG
Q_3	59.23%	65.08%	68.34%	71.70%	63.92%	72.76%
Q_H^{obs}	66.34%	66.73%	78.28%	79.13%	67.56%	74.02%
Q_E^{obs}	20.74%	46.32%	42.55%	58.01%	29.54%	56.85%
Q_C^{obs}	72.80%	73.19%	73.14%	72.47%	78.29%	79.75%
Q_H^{pred}	61.87%	68.64%	70.56%	73.91%	69.47%	78.76%
Q_E^{pred}	56.45%	58.88%	69.35%	68.28%	73.50%	71.93%
Q_C^{pred}	57.77%	64.72%	66.21%	71.22%	59.08%	69.00%

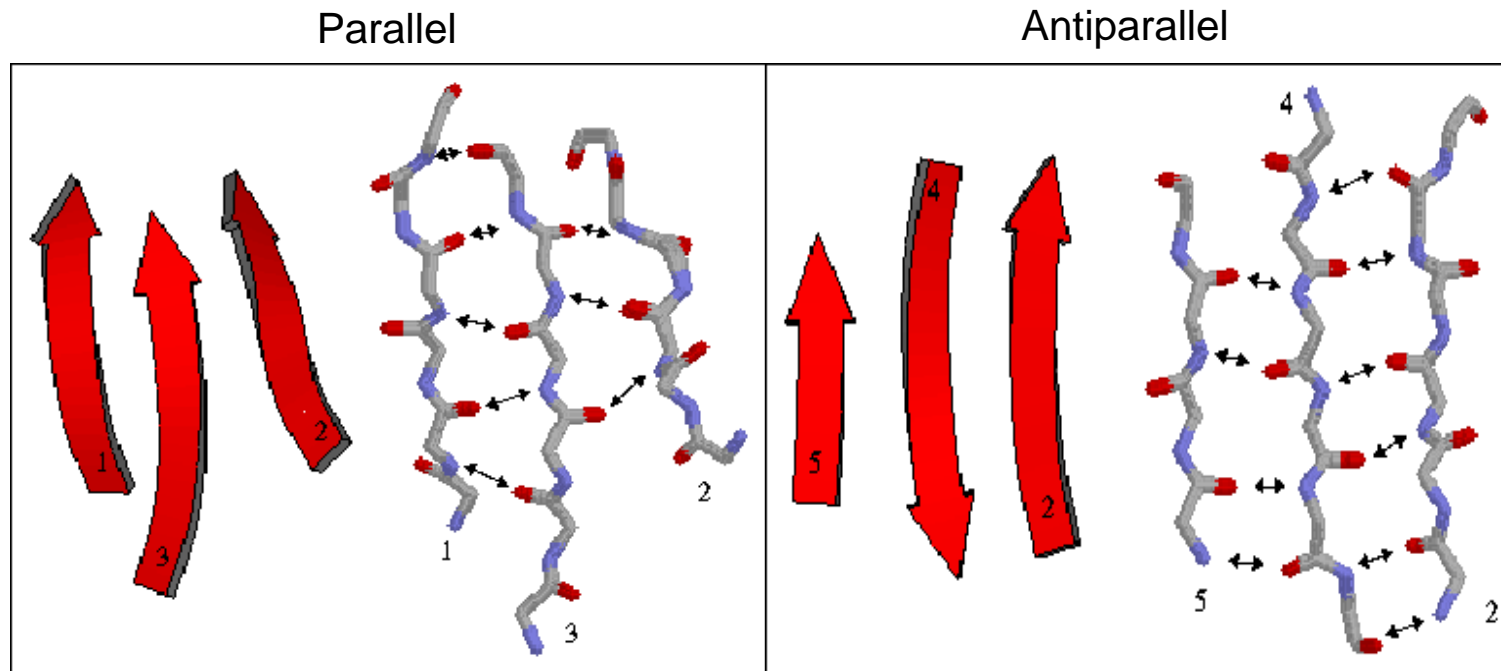
“Sequence Only” denotes the algorithm of Schmidler et al. 2000; “MSAP” denotes our approach using multiple sequence alignment profiles; “PSSM” denotes using position specific score matrices. Q_3 denotes the overall accuracy. $Q^{obs} = \frac{TruePositive}{TruePositive+FalseNegative}$ and $Q^{pred} =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

Predictive Results on the Protein Data of CASP

	CASP2 (20 chains)	CASP3 (36 chains)	CASP4 (40 chains)	CASP5 (56 chains)
Q_3	73.40%	71.12%	74.32%	74.03%
Q_H^{obs}	76.62%	73.12%	80.22%	80.43%
Q_E^{obs}	61.29%	56.35%	57.81%	59.52%
Q_C^{obs}	77.73%	78.88%	78.00%	76.81%
Q_H^{pred}	79.71%	74.91%	81.33%	76.95%
Q_E^{pred}	76.48%	78.39%	76.19%	78.10%
Q_C^{pred}	67.36%	65.99%	67.28%	69.88%

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 $Q^{pred} = \frac{TruePositive}{TruePositive+FalsePositive}$.

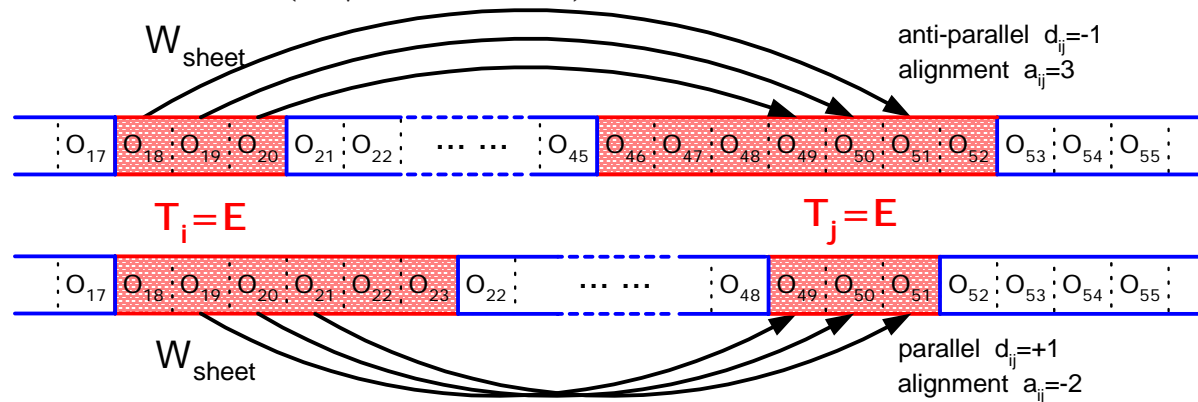
Long-range Interactions in β -sheets



The β -sheet space is the set of all the possible combinations of β -sheets;
 A set of interaction variables, \mathcal{I} , to describe one possible case.

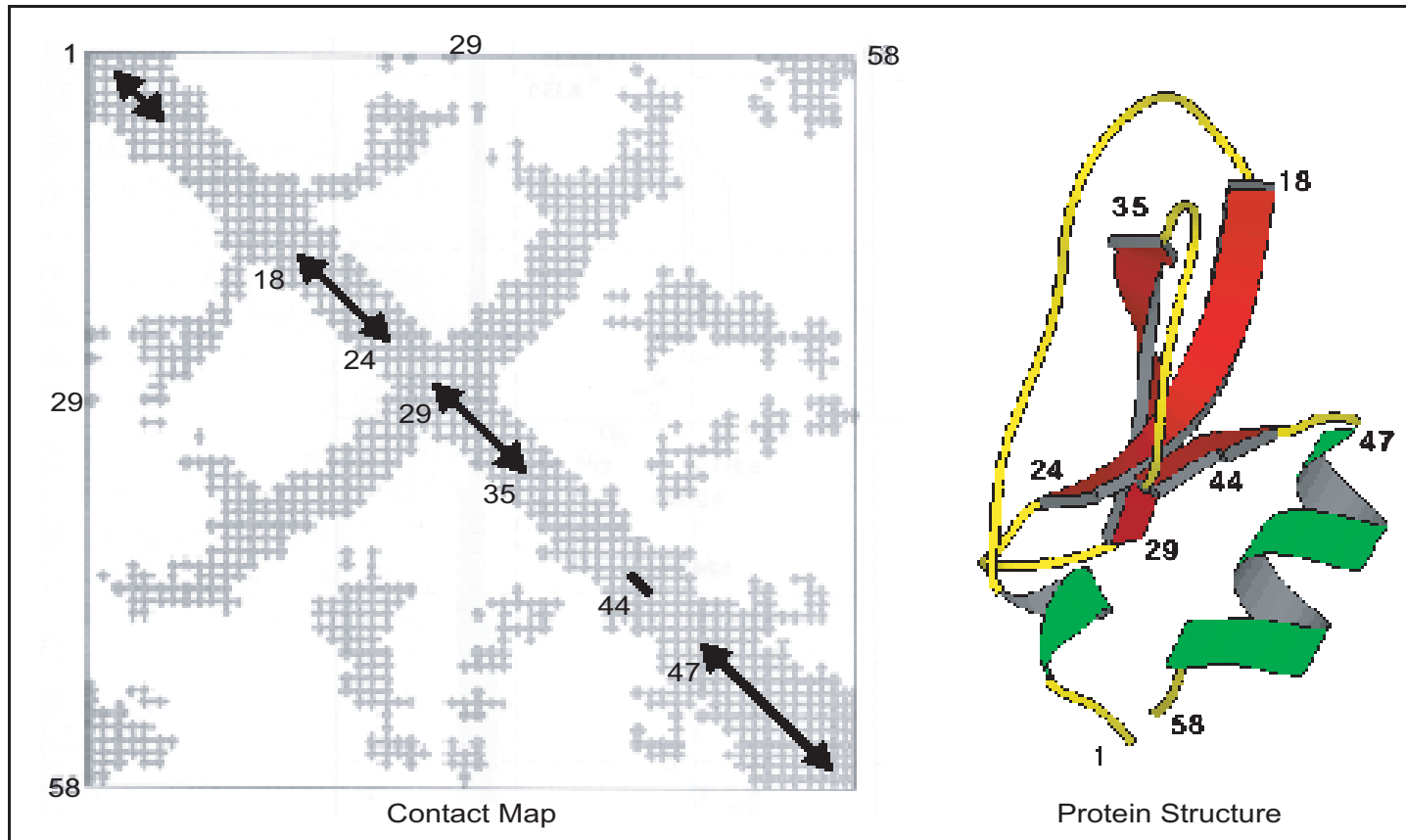
Enhanced Framework

- Prior: $\mathcal{P}(m, e, T, \mathcal{I}) = \mathcal{P}(\mathcal{I}|m, e, T)\mathcal{P}(m, e, T)$, where $\mathcal{P}(\mathcal{I}|m, e, T)$ is the distribution over β -sheet space;
- Likelihood: $\mathcal{P}(O|m, e, T, \mathcal{I})$



- Posterior: $\mathcal{P}(m, e, T, \mathcal{I}|O) \propto \mathcal{P}(O|m, e, T, \mathcal{I})\mathcal{P}(m, e, T, \mathcal{I})$;
- **Metropolis-Hasting scheme with reversible-jumps** (Green, 1995) can be used to collect samples in $\mathcal{P}(m, e, T|O) = \sum_{\mathcal{I}} \mathcal{P}(m, e, T, \mathcal{I}|O)$.

Contact Maps



Inference on Contact Maps

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

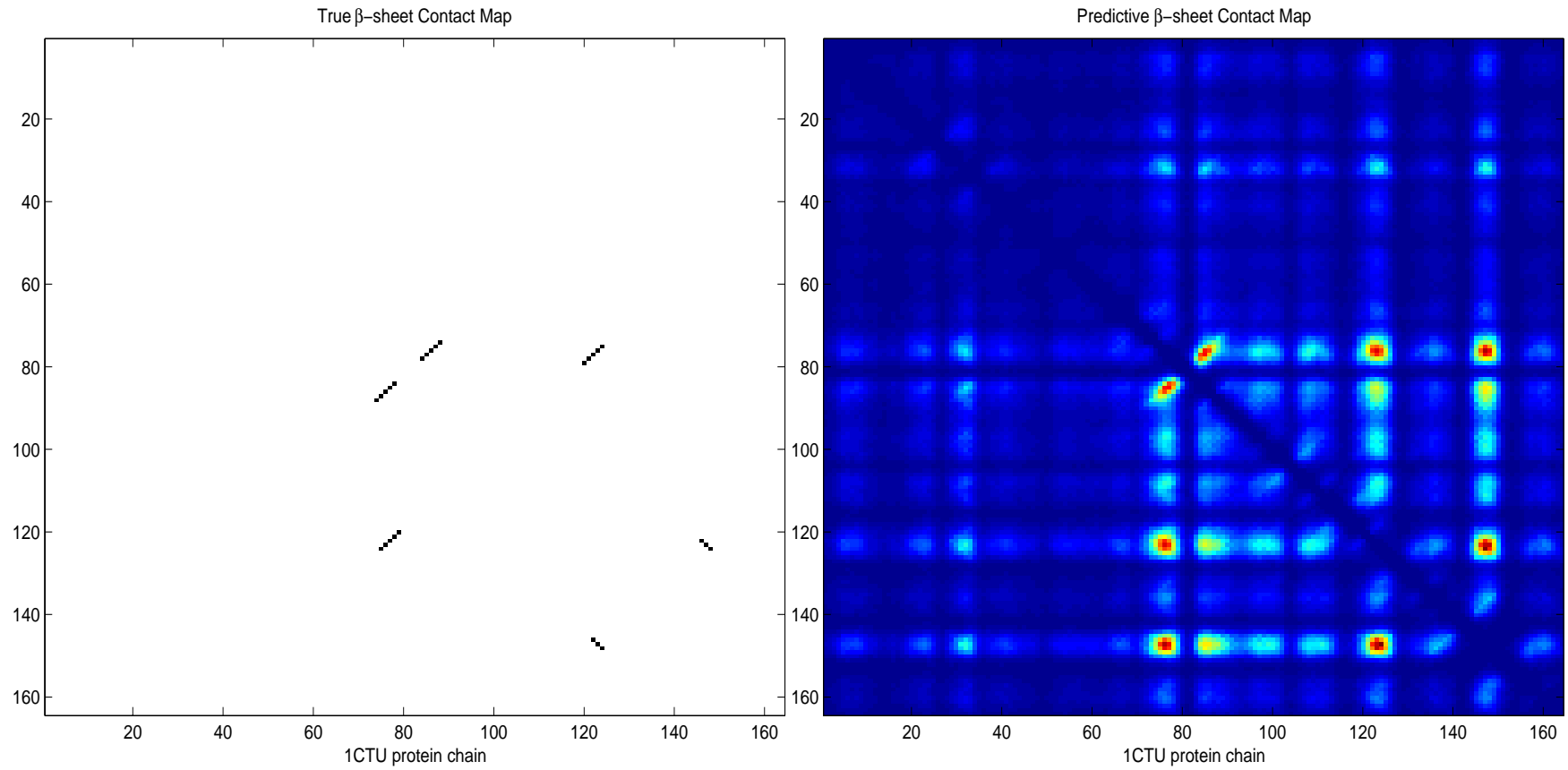
$$\mathcal{C}^{ij}(\mathcal{I}) = \begin{cases} 1 & \text{if } O_i \text{ and } O_j \text{ are paired in the interaction set } \mathcal{I}; \\ 0 & \text{otherwise} \end{cases}$$

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

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

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

An Example



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);