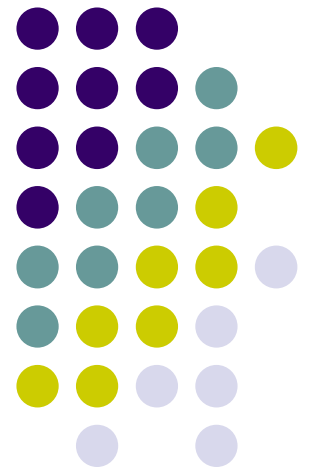


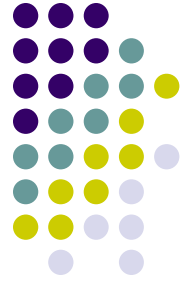
Bayesian Protein Structure Prediction

Authors: Scott C. Schmidler, Jun S. Liu, Douglas L. Brutlag

Presented By:
Srividhya Rajendran

For CSE-6392
Reasoning with Uncertainty





Introduction to Proteins

What are Proteins??

Proteins are polymers formed by the linkage of amino acids via a peptide bond.

Protein Structure:

1. Primary Structure
2. Secondary Structure
3. Tertiary Structure

Primary Structure:

Refers to the specific "linear" sequence of amino acids bonded together by peptide bonds also known as a polypeptide chain.

Introduction to Proteins



Secondary Structure:

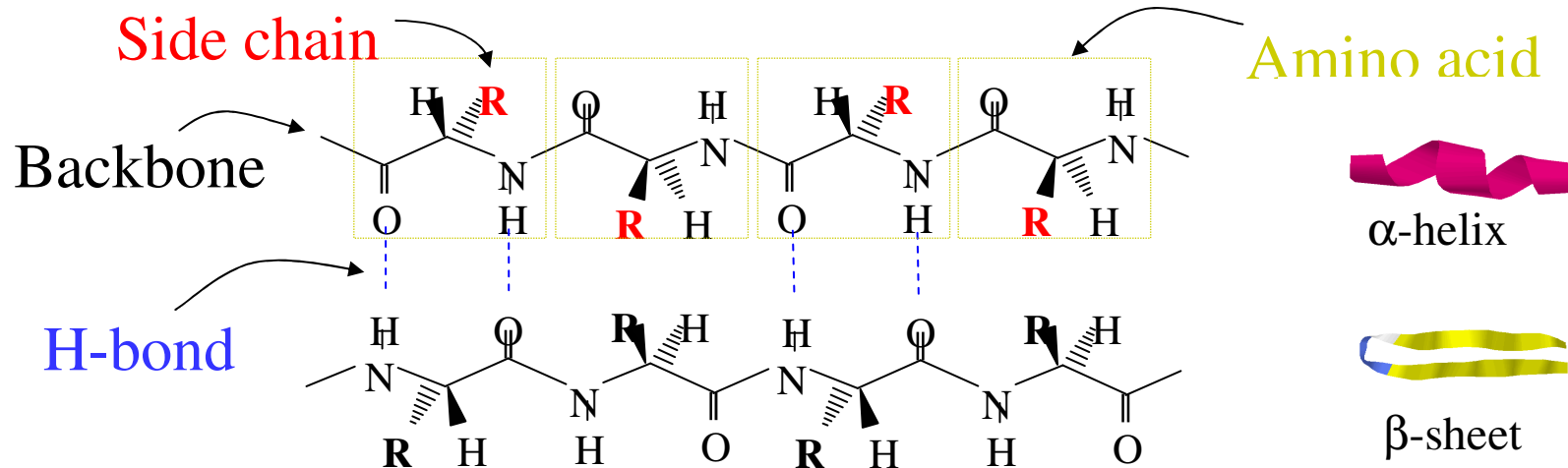
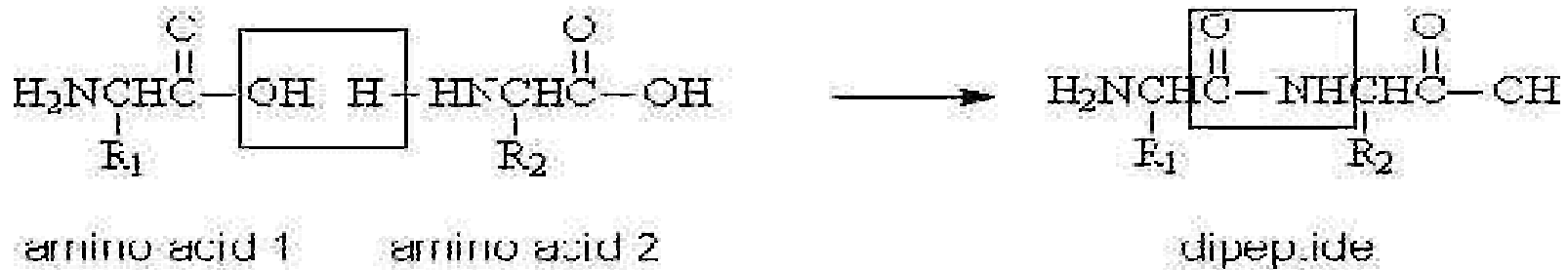
is "local" ordered structure brought about via hydrogen bonding mainly within the peptide backbone. (stabilizes the primary structure).

1. α Helix
2. β Sheets
3. Hairpin Loops /Reverse Turns.

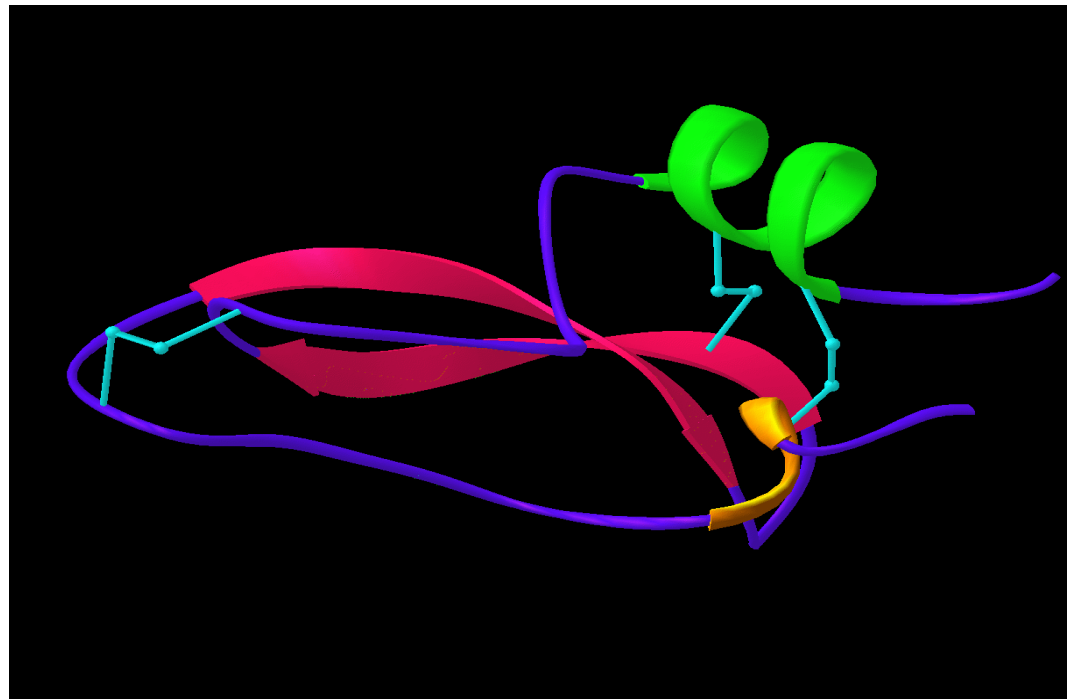
Tertiary Structure:

It is the compact 3-dimensional shape formed by folding of secondary structures of proteins (by a process know as protein folding).

1-D and 2-D Protein Structure



3-D Protein Structure



Bovine Pancreatic Trypsin Inhibitor

- a) β -Sheets are colored in **Hot Pink** color, b) α - helix is colored in **green** and **golden** color

(source: <http://www-nmr.cabm.rutgers.edu/photogallery>)

Protein Structure Prediction



To determine proteins:

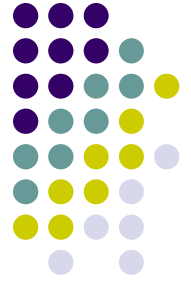
1. Secondary and
2. Tertiary Structures

The way the protein fold themselves define the role they play in life.

Problem in Prediction:

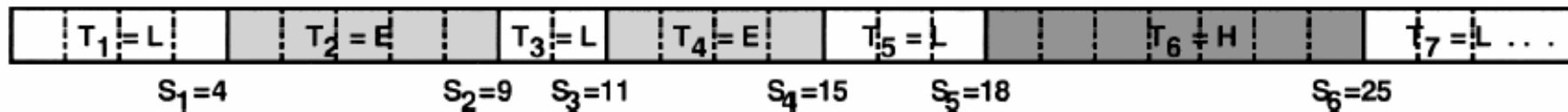
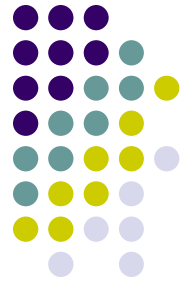
- Human genome project contains approx 30,000-100,000 genes.
- Protein folding process is very poorly understood.

Importance of Protein structure Prediction



- The shape of a protein determines its function.
- Knowledge of structure is used in many ways:
 - Drug design
 - Design of synthetic proteins
 - Re-engineering defective proteins
(may be a cure for disease like Alzheimer's or Fatal familial insomnia)
- Genome projects are providing sequences for many proteins whose structure will need to be determined

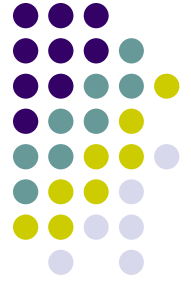
Secondary Structure Prediction



Representation of the secondary structure of a protein sequence in terms of structural *segments*. The parameters shown represent the segment types $T = (L, E, L, E, L, H, L, \dots)$ and endpoints $S = (4, 9, 11, 15, 18, 25, \dots)$.

1. $R = (R_1, R_2, \dots, R_n)$ be sequence of n amino acid Residue
2. $S = \{ i: \text{Struct}(R_i) \neq \text{Struct}(R_{i+1}) \}$ sequence of m positions representing the end of individual structural segment.
3. $T = (T_1, T_2, T_3, \dots, T_m)$ is sequence of secondary structural types for each representative segment.
4. $\forall i T_i \in \{H, E, L\}$ where $H - \alpha$ Helix, $E - \beta$ Strands, $L -$ Coils/Loops.

Prediction of secondary structure requires to get values of m , $S = (S_1, S_2, S_3, \dots, S_m)$ and $T = (T_1, T_2, T_3, \dots, T_m)$ corresponding to a known amino acid sequence $R = (R_1, R_2, \dots, R_n)$ i.e. $P(m, S, T | R)$.



Secondary Structure Prediction

To compute $P(m, S, T | R)$ using Bayesian approach

We start with joint probability distribution $P(R, m, S, T)$ which may be factored by conditional independence of inter-segment residues.

1.

$$P(R | m, S, T) = \prod_{j=1}^m P(R_{[S_{j-1}+1:S_j]} | S, T)$$

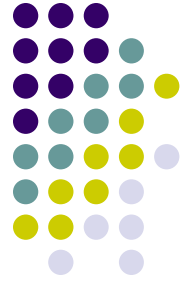
Given the above equation we only need $P(m, S, T)$ to completely satisfy the joint probability distribution.

2.

$$P(m, S, T) = P(m) \prod_{j=1}^m P(T_j | T_{j-1}) P(S_j | S_{j-1}, T_j)$$

Here it is assumed that each segment type only depends on its nearest neighbors and the conditioning of S_j on (S_{j-1}, T_j) allows for explicit modeling of different length distributions of each segment type observed in Protein Data Bank.

Secondary Structure Prediction



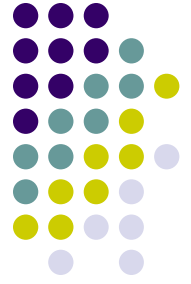
Equation 1. can be further expanded for each segment type.

Helical Model:

$$\begin{aligned} P(R_{[S_{j-1}+1:S_j]} | S_{j-1}, S_j, H) &= \prod_{i=S_{j-1}+1}^{S_{j-1}+\ell_N^H} P_{N_{i-S_{j-1}}}^H(R_i | R_{[S_{j-1}+1:i-1]}) \\ &\times \prod_{i=S_{j-1}+\ell_N^H+1}^{S_j-\ell_C^H} P_I^H(R_i | R_{[S_{j-1}+1:i-1]}) \\ &\times \prod_{i=S_j-\ell_C^H+1}^{S_j} P_{C_{S_j-i+1}}^H(R_i | R_{[S_{j-1}+1:i-1]}). \end{aligned}$$

Where ℓ_n^H indicates the length of helix N-Cap model, N_i , C_i indicates the i th position from the N- and C- termini respectively. and I indicates an Internal (noncap) position.

Secondary Structure Prediction



$$\alpha(j, t) = \sum_{v=l}^{j-1} \sum_{l \in SS} \alpha(v, l) P(R_{[v+1:j]} | S_{prev} = v, S = j, T = t, \theta) \\ \times P(S = j | T = t, S_{prev} = v, \theta) P(T = t | T_{prev} = l, \theta)$$

$$\beta(j, t) = \sum_{v=j+1}^n \sum_{l \in SS} \beta(v, l) P(R_{[j+1:v]} | S_{next} = v, S = j, T_{next} = l, \theta) \\ \times P(S_{next} = v | S = j, T_{next} = 1, \theta) P(T_{next} = l | T = t, \theta)$$

Where SS= {H, E, L} set of possible structural types, θ model parameters.



Secondary Structure Prediction

Equation 1. can be further expanded for each segment type.

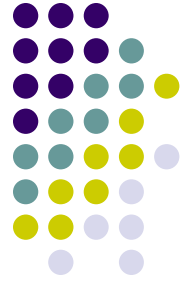
Helical Model:



$$P(R_{[S_{j-1}+1:S_j]} | S_{j-1}, S_j, H) =$$

$$P_{N_0}^H(N)P_{N_1}^H(L|N)P_{N_2}^H(A|N,L)P_{N_3}^H(K|N,L,A) \times P_I^H(M|N,\dots,K)P_I^H(V|N,\dots,M)\dots P_I^H(A|N,\dots,K)$$

$$\times P_{C_3}^H(I|N,\dots,A)P_{C_2}^H(L|N,\dots,I)P_{C_1}^H(K|N,\dots,L)P_{C_0}^H(D|N,\dots,K)$$



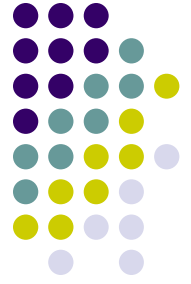
$$\alpha(j, t) = \sum_{v=l}^{j-1} \sum_{l \in SS} \alpha(v, l) P(R_{[v+1:j]} | S_{prev} = v, S = j, T = t, \theta) \\ \times P(S = j | T = t, S_{prev} = v, \theta) P(T = t | T_{prev} = l, \theta)$$

$$\beta(j, t) = \sum_{v=j+1}^n \sum_{l \in SS} \beta(v, l) P(R_{[j+1:v]} | S_{next} = v, S = j, T_{next} = l, \theta) \\ \times P(S_{next} = v | S = j, T_{next} = 1, \theta) P(T_{next} = l | T = t, \theta)$$

$$P(T_{R_i} = t | R, \theta) = \sum_{j=i-D+1}^{i-1} \sum_{k=i}^{j+D-1} \sum_{l \in SS} \alpha(j, l) \beta(k, t) P(R_{[j+1:k]} | S_{prev} = j, S = k, T = t, \theta) \\ \times P(S = k | S_{prev} = j, T = t, \theta) P(T = t | T_{prev} = l, \theta) / Z$$

Where $P(T_{R[i]} | R, \theta)$ is the marginal posterior distribution over structural types at position i . Z is normalizing constant.

Non local effects on Secondary Structure Prediction



All the above Prediction of secondary sequence assumed that there no importance of non local contacts. i.e. amino acids which are sequentially distant in primary structure.

The non local contacts are formed when protein sequence folds back on itself in 3-D.

Importance of these non local contacts is still under debate.



Inter-Segment Interactions

All the secondary structure prediction done until are violated if non locals contacts are considered to be important for secondary structure prediction.

Joint Segment Likelihoods: For 2 segments j and k interacting we replace the terms

$$P(R_{[S_{j-1}+1:S_j]} | S_{j-1}, S_j, T_j) \text{ and } P(R_{[S_{k-1}+1:S_k]} | S_{k-1}, S_k, T_k)$$

With $P(R, m, S, T, \mathcal{P}) \propto P(m, S, T, \mathcal{P}) \prod_{j \notin \mathcal{P}} P(R_{[S_{j-1}+1:S_j]} | S, T, m, \mathcal{P}) \times$

$$\prod_{(j,k) \in \mathcal{P}} P(R_{[S_{j-1}+1:S_j]}, R_{[S_{k-1}+1:S_k]} | S, T, m, \mathcal{P})$$

where \mathcal{P} is the set of pairs interacting segments.

However computation of posterior quantities now involves maximization/marginalization over all possible segment interactions and thus is intractable computation.

MCMC Segmentation

(Markov chain Monte Carlo Segmentation)



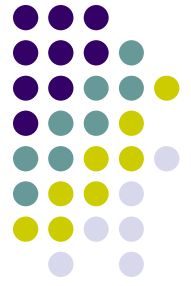
This allows us to approximate inference in the Models. Further a reversible jump approach is used to take care of varying dimensionality (m and P).

To construct a Markov chain on the space of sequence segmentations, we define the following set of Metropolis proposals:

- *Type switching:*
Given a segmentation (m, S, T) , propose a move to segmentation (m, S, T^*) where $T_j^* = T_j, j \neq k$ for some k chosen uniformly at random or by systematic scan, and $T_k^* \sim U[\{H, E, L\}]$.
- *Position change:*
Given (m, S, T) , propose (m, S^*, T) with $S_j^* = S_j, j \neq k$ for some k and $S_k^* \sim U[S_{k-1} + 1, S_{k+1} - 1]$.

MCMC Segmentation

(Markov chain Monte Carlo Segmentation)



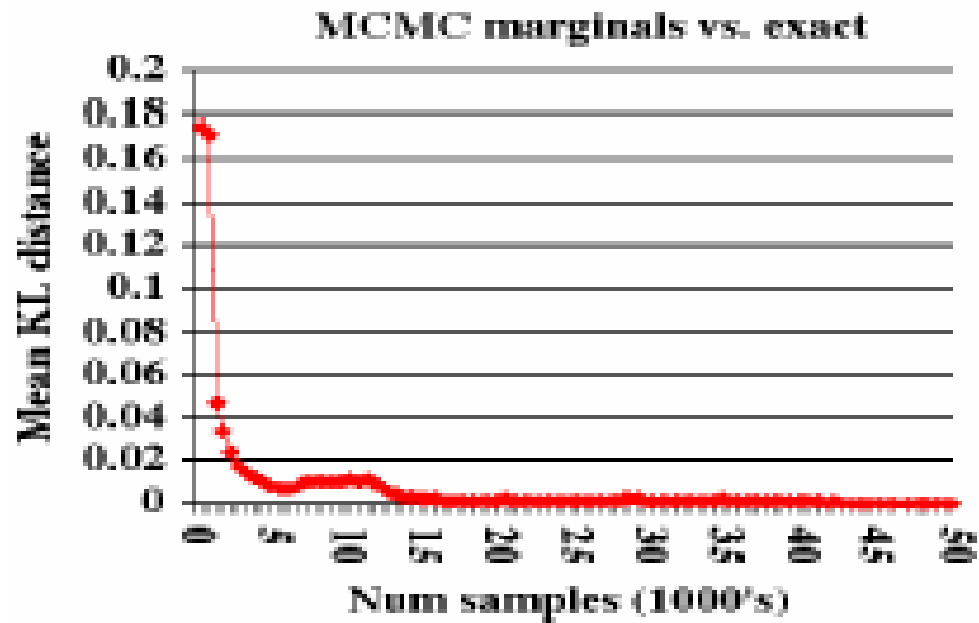
- *Segment split:*
Given (m, S, T) , propose (m^*, S^*, T^*) with $m^* = m + 1$ segments by splitting segment $1 \leq k \leq m$ into two new segments $(k^*, k^* + 1)$ where $k \sim U[1, m]$, $S_{k^*+1}^* = S_k$, and $S_{k^*}^* \sim U[S_{k-1} + 1, S_k - 1]$. With probability $\frac{1}{2}$, we set $T_{k^*} = T_k$ and $T_{k^*+1} = T_{new}$ with T_{new} chosen uniformly, and with probability $\frac{1}{2}$ do the reverse.
- *Segment merge:*
Similar to *segment split*, but a randomly chosen segment is merged into a neighbor and $m^* = m - 1$.

For joint segment models such as (4.2), additional proposal moves must be added involving interacting segments:

- *Segment join:*
Proposes a replacement of two non-interacting segments (S_j, T_j) and (S_k, T_k) , $(j, k) \notin \mathcal{P}$ with an interaction (S_j, S_k, T_j, T_k) , $(j, k) \in \mathcal{P}$. In Section 5 below, this corresponds to replacing two independent β -strands with a β -sheet consisting of the two strands joined.



Results

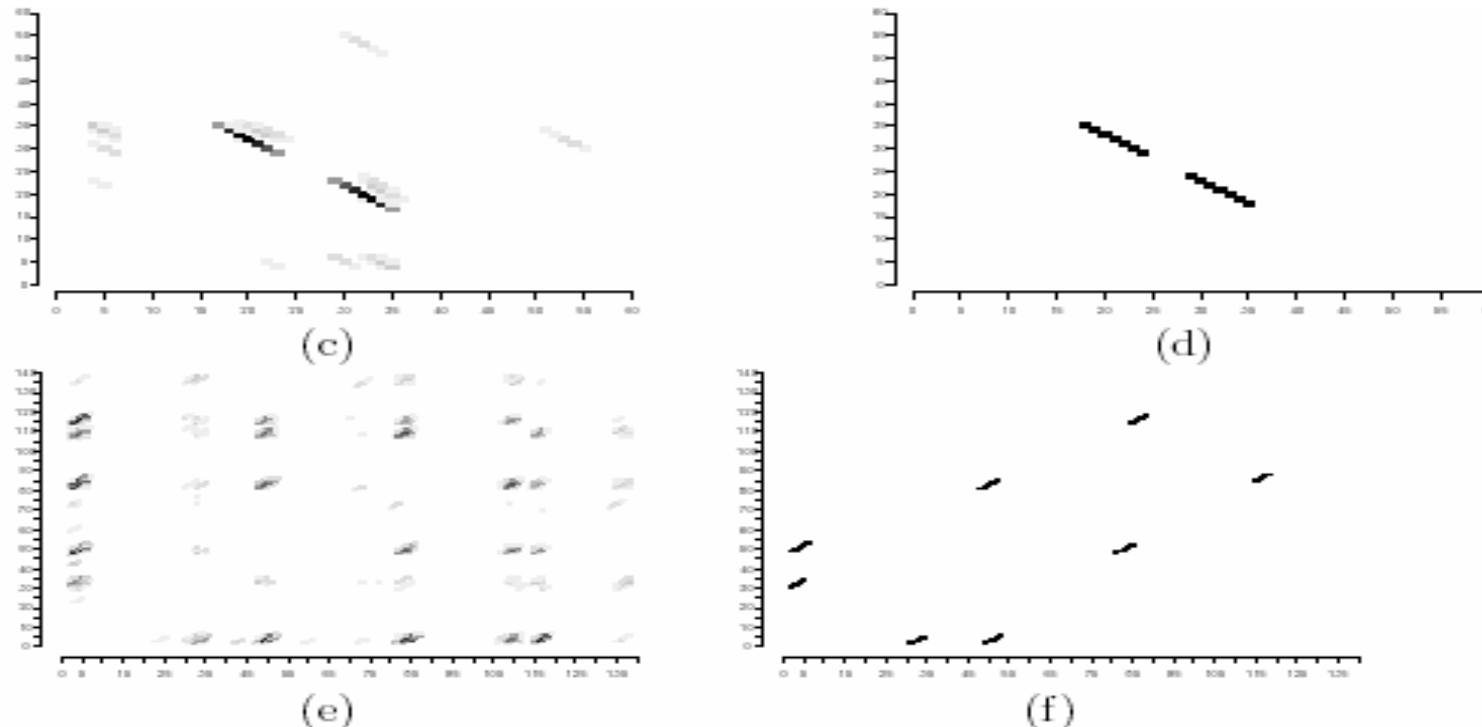


(a)

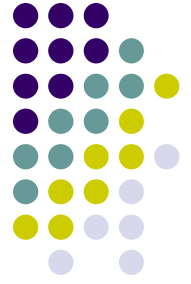
Shows convergence of MCMC simulations to exact calculations. Plot is a mean Kullback-Leibler divergence between mean marginal distributions obtained from exact and MCMC calculations.



Results



$KL(p, q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$. (b) True structure of bovine pancreatic trypsin inhibitor (BPTI). (c) Predicted and (d) true β -strand contacts for BPTI. Axes are sequence position, and shading of (x, y) is proportional to predicted probability of contact for positions x, y . The β -hairpin contacts are predicted with high probability. The *maximum a posteriori* sheet topology correctly identifies β -strand locations and register (not shown). Pairings representing register shifts are also observed with lower probability. (e) Predicted and (f) true contacts for flavodoxin, showing significant uncertainty in correct pairing of strand segments.



References:

1. Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77(2), 257–286.
2. Schmidler, S. C., Liu, J. S., and Brutlag, D. L. (2000). Bayesian segmentation of protein secondary structure. *J. Comp. Biol.*, 7(1):233-248.
3. Schmidler, S. C., Liu, J. S., and Brutlag, D. L. (2001). Bayesian Protein Structure Prediction, *Case Studies in Bayesian Statistics, Vol. 5*, pp 363-378.
4. Dr. Frank R. Gorga, Department of Chemical Science, Bridgewater State College, Bridgewater, MA, Introduction to Protein Structure. (<http://webhost.bridgew.edu/fgorga/proteins/default.htm>)
5. Image of bovine Pancreatic trypsin Inhibitor. (<http://www-nmr.cabm.rutgers.edu/photogallery/>)
6. Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.
7. Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711-32.

THANK YOU



QUESTIONS

