Bayesian Protein Structure Prediction

Scott C. Schmidler
Jun S. Liu
Douglas L. Brutlag

ABSTRACT An important role for statisticians in the age of the Human Genome Project has developed in the emerging area of "structural bioinformatics". Sequence analysis and structure prediction for biopolymers is a crucial step on the path to turning newly sequenced genomic data into biologically and pharmaceutically relevant information in support of molecular medicine. We describe our work on Bayesian models for prediction of protein structure from sequence, based on analysis of a database of experimentally determined protein structures. We have previously developed segment-based models of protein secondary structure which capture fundamental aspects of the protein folding process. These models provide predictive performance at the level of the best available methods in the field (Schmidler et al., 2000). Here we show that this Bayesian framework is naturally generalized to incorporate information based on non-local sequence interactions. We demonstrate this idea by presenting a simple model for β-strand pairing and a Markov chain Monte Carlo (MCMC) algorithm for inference. We apply the approach to prediction of 3-dimensional contacts for two example proteins.

1 Introduction

The Human Genome Project estimates that sequencing of the entire complement of human DNA will be completed in the year 2003, if not sooner (Collins et al., 1998). At the same time a number of complete genomes for pathogenic organisms are already available, with many more under way. Widespread availability of this data promises to revolutionize areas of biology and medicine, providing fundamental insights into the molecular mechanisms of disease and pointing the way to the development of novel therapeutic agents. Before this promise can be fulfilled however, a number of significant hurdles remain. Each individual gene must be located within the 3 billion bases of the human genome, and the functional role of its associated protein product identified. This process of functional characterization, and subsequent development of pharmaceutical agents to affect that function, is greatly aided by knowledge of the 3-dimensional structure into which the protein folds. While the sequence of a protein can be determined directly from the DNA of the gene which encodes it, prediction of the 3-dimensional structure of the protein from that sequence remains one of the great open problems of science. Moreover, the scale of the problem (the human genome is projected to contain approximately 30,000-100,000 genes) necessitates the development of *computational* solutions which capitalize on the laboriously acquired experimental structure data. The field of research which has sprung up in support of these efforts is coming to be known as "structural bioinformatics", and poses a number of scientifically important and theoretically challenging problems involving data analysis and prediction. Emerging efforts to develop and make publicly available a large, structurally diverse set of experimental data as a structural analog of the Human Genome Project (Burley et al., 1999; Montelione and Anderson, 1999) promise to provide a multitude of statistical problems within this emerging research area.

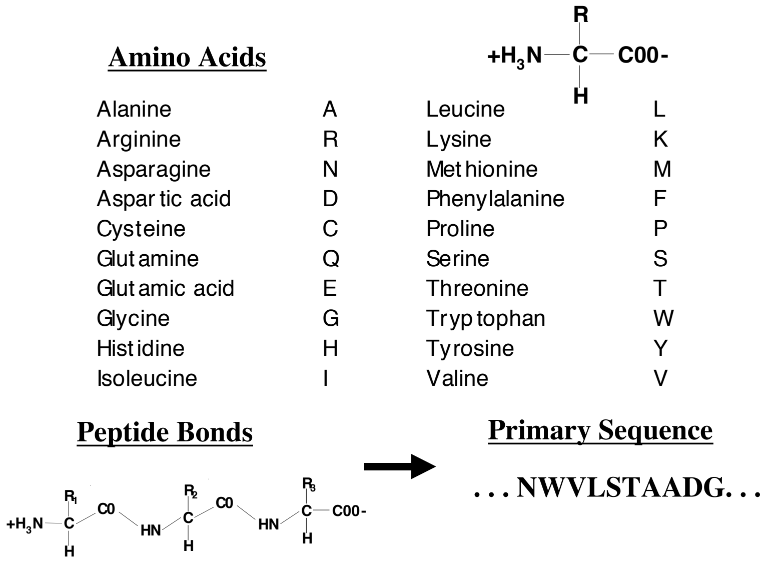
2 Protein Structure Prediction

2.1 Proteins and their structures

A protein sequence is a linear heteropolymer, meaning simply that it is an unbranched chain of molecules with each "link" in the chain made up by one of the twenty *amino acids* (see Figure 1). Proteins perform the vast majority of the biochemistry required by living organisms, playing various catalytic, structural, regulatory, and signaling roles required for cellular development, differentiation, replication, and survival. The key to the wide variety of functions exhibited by individual proteins is not the linear sequence as shown in Figure 1 however, but the three dimensional configuration adopted by this sequence in its native environment. In order to understand protein function at the molecular level then, it is crucial to study the structure adopted by a particular sequence. Unfortunately the physical process by which a sequence achieves this structure, known as *protein folding*, remains poorly understood despite decades of study. In particular, serious difficulties present themselves when one attempts to *predict* the folded structure of a given protein sequence.

2.2 Protein structure prediction

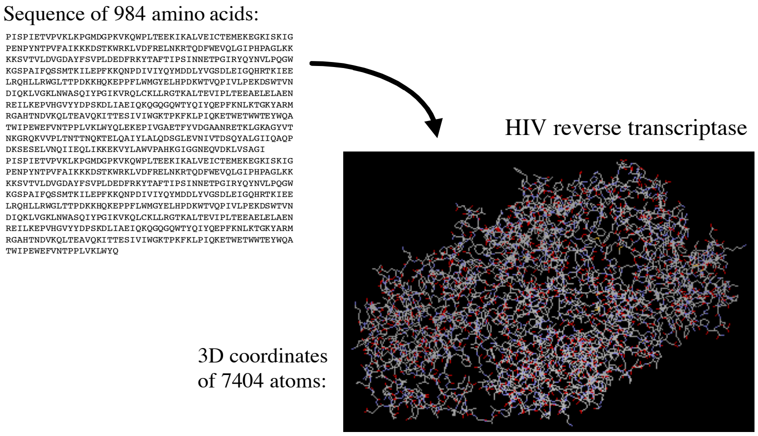
The basic problem of protein structure prediction is summarized in Figure 2. The goal is to take an amino acid sequence, represented as a sequence of letters as shown in Figure 1, and predict the three dimensional conformation adopted by the protein in its native (folded) state. The difficulties in doing so are numerous, and significant effort has been directed towards developing approximate methods based on reduced representations



Amino Acids

Alanine	A	Leucine	L
Arginine	R	Lysine	K
Asparagine	N	Met hionine	M
Aspar tic acid	D	Phenylalanine	F
Cysteine	C	Proline	P
Glut amine	Q	Serine	S
Glut amic acid	E	Threonine	T
Glycine	G	Tryp tophan	W
Hist idine	H	Tyrosine	Y
Isoleucine	I	Valine	V

FIGURE 1. The basic components of protein structure. Proteins are made up of twenty naturally occurring amino acids linked by peptide bonds to form linear polymers. Each amino acid is represented by a letter of the alphabet to produce a protein sequence.



Sequence of 984 amino acids:

```
PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIG
PENPYNTPVFAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGIPHPAGLKK
KKSVTVLDVGDAYFSVPLDEDFRKYTAFTIPSINNETPGIRYQYNVLPQGW
KGSPAIFQSSMTKILEPFKKQNPDIVIYQYMDDLYVGSDLEIGQHRTKIEE
LRQHLLRWGLTTPDKKHQKEPPFLWMGYELHPDKWTVQPIVLPEKDSWTVN
DIQKLVGKLNWASQIYPGIKVRQLCKLLRGTKALTEVIPLTEEAELELAEN
REILKEPVHGVYYDPSKDLIAEIQKQGQGQWTYQIYQEPFKNLKTGKYARM
RGAHTNDVKQLTEAVQKITTESIVIWGKTPKFKLPIQKETWETWWTEYWQA
TWIPEWEFVNTPPLVKLWYQLEKEPIVGAETFYVDGAANRETKLGKAGYVT
NKGRQKVVPLTNTTNQKTELQAIYLALQDSGLEVNIVTDSQYALGIIQAQP
DKSESELVNQIIEQLIKKEKVYLAWVPAHKGIGGNEQVDKLVSAGI
PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIG
PENPYNTPVFAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGIPHPAGLKK
KKSVTVLDVGDAYFSVPLDEDFRKYTAFTIPSINNETPGIRYQYNVLPQGW
KGSPAIFQSSMTKILEPFKKQNPDIVIYQYMDDLYVGSDLEIGQHRTKIEE
LRQHLLRWGLTTPDKKHQKEPPFLWMGYELHPDKWTVQPIVLPEKDSWTVN
DIQKLVGKLNWASQIYPGIKVKQLCKLLRGTKALTEVIPLTEEAELELAEN
REILKEPVHGVYYDPSKDLIAEIQKQGQGQWTYQIYQEPFKNLKTGKYARM
RGAHTNDVKQLTEAVQKITTESIVIWGKTPKFKLPIQKETWETWWTEYWQA
TWIPEWEFVNTPPLVKLWYQ
```

FIGURE 2. The protein structure prediction problem: predicting the 3D coordinates of a folded protein from the amino acid sequence. The example protein shown is HIV reverse transcriptase, a DNA polymerase required for HIV replication and therefore a target for pharmaceutical development.

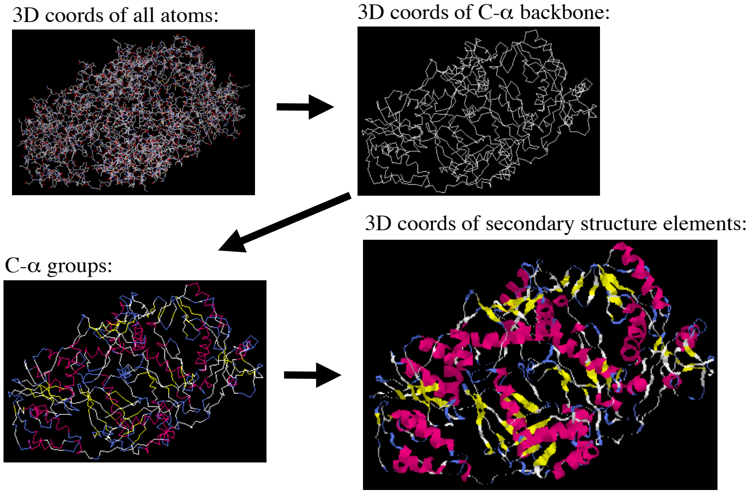

3D coords of all atoms:
3D coords of C-α backbone:
C-α groups:
3D coords of secondary structure elements:

FIGURE 3. Successive abstraction of the problem: From atomic coordinates to α-carbon backbone to segments of secondary structure.

of proteins (see Neumaier (1997) for a review from a mathematical perspective). Here we focus only on one such abstraction of the problem, which characterizes a protein structure by short segments of regular repeated conformation, known as *secondary structure*. Figure 3 shows the process of successive abstraction leading to a representation of protein structure in terms of secondary structure vectors in space. The secondary structure elements of greatest interest are helical regions known as *α-helices*, and extended regions known as *β-strands*, which join together to form *β-sheets*. Figure 4 shows both an α-helix and a β-sheet. The *secondary structure prediction problem* is the task of predicting the location of α-helices and β-strands in an amino acid sequence, in the absence of any knowledge of the tertiary structure of the protein. The task is thus to predict a 1-dimensional summary of the 3-dimensional folded structure, as shown in Figure 4. This 1D summary is typically formulated as a 3-state problem, with all positions classified as being in either α-helix (H), extended β-strand (E), or loop/coil (L) conformation. Accurate secondary structure predictions are of considerable interest, because knowledge of the location of secondary structure elements can be used for approximate folding algorithms (Monge et al., 1994; Eyrich et al., 1999) or to improve fold recognition algorithms (Fischer and Eisenberg, 1996; Russell et al., 1996), which can in many cases yield low-resolution 3D structures for the folded protein. Because of this, secondary structure prediction has received a great deal of attention over several decades, but remains a difficult problem (see (Barton, 1995) or references in (King and Sternberg, 1996; Schmidler et al., 2000) for a review). Standard approaches predict each sequence position inde-

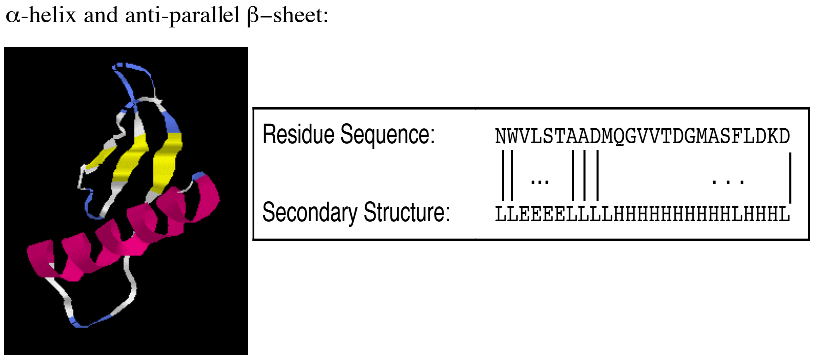

α-helix and anti-parallel β–sheet:
Residue Sequence:
NWVLSTAADMQGVVTDGMASFLDKD
...
. . .
Secondary Structure:
LLEEEELLLLHHHHHHHHHHLHHHL

FIGURE 4. The secondary structure of a protein is defined by the local backbone conformation at each position. Secondary structure elements of greatest interest include α-helices (dark) and extended β-strands which come together to form β-sheets (light). These are represented as H and E respectively in the 1D summary. Remaining positions are represented by L for loop/coil.

pendently based on a local surrounding subsequence. The most accurate such “window-based” methods currently use neural-networks or nearest-neighbor classifiers. A widely recognized drawback of these approaches is lack of interpretability of model parameters, yielding little insight into the important factors in protein folding.

2.3 Non-local effects

One of the difficulties in predicting secondary structure at high accuracy is the importance of *non-local* contacts in protein folding. Amino acids which are sequentially distant in the primary structure may be in close physical proximity in the tertiary structure, as the sequence folds back on itself in three dimensions. The relative importance of local vs. non-local effects in determining protein folds is still under debate (Baldwin and Rose, 1999; Dill, 1999), but it is clear that non-local effects can be important. For example, identical 5- and 6- amino acid subsequences have been located which take on different local conformations in different proteins (Kabsch and Sander, 1984; Cohen et al., 1993). Moreover, an 11 amino acid “chameleon” sequence has been designed which folds into an α-helical conformation when placed at one position in a particular protein, and a β-strand conformation when placed at a different position of the same protein (Minor and Kim, 1996). A possible explanation for such observations is the effect of non-local contacts in determining local structure.

Regardless of their importance for driving the physical folding process, non-local interactions induce correlations in the sequence which can provide useful information for protein structure prediction. For example, the side chains of adjacent β-strands in a β-sheet will experience a similar chemical environment, and therefore acceptable mutations in these strands will exhibit correlations (Lifson and Sander, 1980; Hutchinson et al., 1998). In general, positions which are in close physical proximity in the tertiary structure may be expected to exhibit correlated mutations, irrespective of their relative positions in sequence. In Section 4, we show how such information can be captured formally for use in prediction.

3 Bayesian Sequence Segmentation

We have developed a Bayesian framework for prediction of protein secondary structure from sequence. Our approach is based on the parameterization of protein sequence/structure relationships in terms of structural *segments*. An overview of the class of models developed is provided here; more details and relations to other statistical models can be found in (Schmidler, 2000).

Let $R = (R_1, R_2, \ldots, R_n)$ be a sequence of n amino acid residues, $S = \{i \mid Struct(R_i) \neq Struct(R_{i+1})\}$ be the positions denoting the ends of m structural segments, and $T = (T_1, T_2, \ldots, T_m)$ be the secondary structure types for the segments. We refer to the set (m, S, T) as a *segmentation* of the sequence R. A segmentation defines an assignment of secondary structure to the sequence R, and we wish to infer the unobserved structure (m, S, T) for an observed sequence R.

We define a joint distribution over (R, m, S, T) of the form:

$$P(R, m, S, T) \propto P(m, S, T) \prod_{j=1}^{m} P(R_{[S_{j-1}+1:S_j]} \mid m, S, T) \qquad (3.1)$$

which factors the joint likelihood $P(R \mid m, S, T)$ by conditional independence of segments given their locations and structural types. Note that marginalization over latent variables (S, T) yields a complex dependency structure among the observed sequence. A special case of (3.1) is a hidden Markov model (HMM).

The segment likelihoods $P(R_{[S_{j-1}+1:S_j]} \mid m, S, T)$ may be of general form. Detailed segment models have been developed to account for experimentally and statistically observed properties of α-helices and β-strands (Schmidler et al., 2000). These models generalize existing stochastic models for secondary structure prediction based on HMMs (Asai et al., 1993; Stultz et al., 1993) in several important ways. The factorization in terms of segments allows modeling of non-independence and non-identity of amino acid distributions at varying positions in the segment. Both position-specific distributions and dependency among positions capture important structural signals such as helix-capping (Aurora and Rose, 1998) and side chain correlations (Klingler and Brutlag, 1994), and these advantages have been explored in detail in previous work.

Given a set of segment likelihoods, we wish to predict the secondary structure for a newly observed protein sequence R. Taking a Bayesian approach, we assign a prior $P(m, S, T)$ and base our predictions on $P(m, S, T \mid R)$, the posterior distribution over secondary structure assignments given the observed sequence. Choice of priors is discussed in Schmidler (2000); one possible approach is to factor $P(m, S, T)$ as a semi-Markov process:

$$P(m, S, T) = P(m) \prod_{j=1}^{m} P(T_j \mid T_{j-1}) P(S_j \mid S_{j-1}, T_j), \tag{3.2}$$

which accounts for empirically observed differences in segment length distributions among structural types.

Under the model defined by (3.1) and (3.2), we consider two possible predictors of interest:

$$\begin{aligned} Struct_{MAP} &= \arg \max_{(m,S,T)} P(m, S, T \mid R, \theta) && (3.3) \\ Struct_{Mode} &= \{\arg \max_{T} P(T_{R_{[i]}} \mid R, \theta)\}_{i=1}^{n} && (3.4) \end{aligned}$$

where $Struct_X$ is a segmentation of R, θ denotes the model parameters, and $P(T_{R_{[i]}} \mid R, \theta)$ is the marginal posterior distribution over structural types at a single position i in the sequence:

$$P(T_{R_{[i]}} \mid R, \theta) = \sum_{(m,S,T)} P(m, S, T \mid R, \theta) \mathbf{1}_{\{T_{R_i} = t\}}$$

(3.3) provides the *maximum a posteriori* segmentation of a sequence, while (3.4) provides the sequence of marginal posterior modes. Note that (3.4) involves marginalization over all possible segmentations. Efficient algorithms have been developed for computation of these estimators under the model defined by (3.1, 3.2) (Schmidler et al., 2000).

4 Incorporation of Inter-Segment Interactions

A fundamental assumption of the class of models described by (3.1) is the conditional independence of amino acids which occur in distinct segments. This assumption enables the exact calculation posterior quantities as mentioned above. However, this assumption is clearly violated in the case of protein sequences, due to the non-local forces involved in protein folding described in Section 2.3. For example, β-sheets consist of β-strands linked by backbone hydrogen bonds (Figure 4). β-sheets are thus a major structural motif which involves interactions of sequentially distant segments to form a stable native fold. Other examples include disulfide bonds and helical bundles. The presence of correlated mutations in such motifs is well known (see Section 2.3). It is often suggested that the inability of window-based prediction algorithms to capture such non-local patterns is responsible for the low accuracy typically achieved in β-strand prediction.

In this section, we extend the framework of Section 3 by introducing joint segment models to account for such inter-segment residue correlations. We describe a MCMC algorithm for inference in this class of models, and demonstrate this approach with a simple model for β-strand pairing in β-sheets.

4.1 Joint segment likelihoods

Modeling of segment interactions may be achieved by definition of *joint* segment likelihoods. For two interacting segments j and k, we replace the terms

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

in the product of (3.1) above with a joint term:

$$P(R_{[S_{j-1}+1:S_j]}, R_{[S_{k-1}+1:S_k]} \mid S_{j-1}, S_j, T_j, S_{k-1}, S_k, T_k) \qquad (4.1)$$

Hence we may include arbitrary joint segment distributions for segment pairs into the model. The extension to three or more segments (as may be required for 4-helix bundles or β-sheets, for example) is obvious. Such models contain pair potentials as a special case; see Schmidler (2000) for a more formal development of this class of models.

Inclusion of terms such as (4.1) leads to a joint distribution of the form:

$$P(R, m, S, T, \mathcal{P}) \propto P(m, S, T, \mathcal{P}) \prod_{j \notin \mathcal{P}} P(R_{[S_{j-1}+1:S_j]} \mid S, T, m, \mathcal{P}) \times \qquad (4.2)$$
$$\prod_{(j,k) \in \mathcal{P}} P(R_{[S_{j-1}+1:S_j]}, R_{[S_{k-1}+1:S_k]} \mid S, T, m, \mathcal{P})$$

where $\mathcal{P}$ is the set of pairs of interacting segments. For example, $\mathcal{P}$ might be the set of β-sheets, with each $p \in \mathcal{P}$ a set of β-strand segments participating in the sheet. Clearly elements $p \in \mathcal{P}$ may include > 2 segments, in which case (4.1) must be defined appropriately. It is also necessary to extend the prior $P(m, S, T)$ to include interactions $P(m, S, T, \mathcal{P})$. For the remainder of this paper we will take $P(m, S, T)$ as defined in (3.2) above, and take $P(\mathcal{P} \mid m, S, T) \propto 1$. This extends the previous semi-Markov prior by a conditionally uniform prior on segment interactions. More realistic priors for $(m, S, T, \mathcal{P})$ are developed in Schmidler (2000).

This joint distribution (4.2) is easily evaluated for any fixed segmentation $(m, S, T, \mathcal{P})$ of a sequence R. However computation of posterior quantities such as (3.3) and (3.4) in the context of (4.2) involves maximization/marginalization over *all possible* segment interactions, an intractable computation.

4.2 Markov chain Monte Carlo segmentation

Despite the difficulty in exact calculation of posterior probabilities, approximate inference in models such as described by (4.2) is feasible using MCMC methods, now a standard tool in the Bayesian statistics community (Gilks et al., 1996). Because the problem has varying dimensionality (m and $\mathcal{P}$ are random variables), we use the reversible jump approach described by (Green, 1995).

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 $S^*_k \sim U[S_{k-1}+1, S_{k+1}-1]$.

- *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 $S^*_{k^*} \sim U\left[S_{k-1}+1, S_k - 1\right]$. 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 $m^* = m - 1$.

All moves are accepted or rejected based on a reversible jump Metropolis criteria (Hastings, 1970; Green, 1995). Together, these steps are sufficient to guarantee ergodicity for models of the form (3.1). The factorization of (3.1) allows Metropolis ratios to be evaluated *locally* with respect to the affected segments. Often the above proposals can be replaced by Gibbs sampling steps which draw from the exact conditional distribution, although it may still be more efficient to *Metropolize* such moves (Liu, 1996).

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.

- *Segment separate*:
  Reverse of *segment join*. For example, splits a 2-strand sheet into two independent strands.

Some care must be taken to realize these proposals for a particular set of joint models, such as those provided in Section 5, especially when interactions may involve more than 2 segments. This is discussed in greater detail by Schmidler (2000), who also provides additional Metropolis moves not required for ergodicity but helpful in improving mixing of the underlying Markov chain.

By defining (4.1) as a product of independent terms and choosing the prior appropriately, we can recover model (3.1) and hence compare this MCMC approach to exact calculations. Figure 5a shows that in this case convergence is quite rapid.

5 Application to Prediction of β-Sheets

As mentioned in Section 2.3, the existence of correlated mutations in β-sheets has been well studied in the protein structure literature. Some attempts have been made to incorporate such long-range sequence correlations into the prediction of protein structure (Hubbard and Park, 1995; Krogh and Riis, 1996; Frishman and Argos, 1996). Here, we show how these interactions are naturally modeled in the Bayesian framework provided by Sections 3 and 4, allowing the information to be formally included in the predictive model.

To demonstrate the application of (4.1) in this case, we define the following joint model for adjacent β-strands to incorporate pairwise side chain correlations:

$$\begin{aligned}
&P(R_{[S_{j-1}+1:S_j]}, R_{[S_{k-1}+1:S_k]} \mid S, T, m, \mathcal{P}) = \\
&\prod_{(h_j,h_k)\in H} P(R_{[S_{j-1}+h_j]}, R_{[S_{k-1}+h_k]} \mid S, T, m, \mathcal{P}) \times \qquad (5.1)\\
&\prod_{h_j \notin H} P(R_{[S_{j-1}+h_j]} \mid S, T, m, \mathcal{P}) \prod_{h_k \notin H} P(R_{[S_{k-1}+h_k]} \mid S, T, m, \mathcal{P})
\end{aligned}$$

where H is the set of (ordered) cross-strand neighboring pairs. This model is simply a product distribution over pairs of neighboring amino acids, the simplest possible model which captures some notion of inter-strand correlation. More detailed models are currently being developed.

This approach has been applied to the prediction of contacts for two test proteins, bovine pancreatic trypsin inhibitor (BPTI) shown in Figure 5b and flavodoxin (not shown). Results for BPTI are shown in Figure 5c,d, where strand pairing is well predicted. Results for flavodoxin are shown in Figure 5e,f, where it is seen that strands are well identified but their interaction pattern has high uncertainty. More accurate interaction models and priors may help resolve this uncertainty. In each case the simulations shown restrict the orientation (parallel vs. anti-parallel) of the interactions to be correct, eliminating a further source of variability. A more extensive evaluation of this approach on a large database is underway, and will be reported elsewhere.

6 Discussion

We have discussed the problem of protein structure prediction, and presented a Bayesian formulation. Models based on factorization of the joint distribution in terms of structural segments naturally capture important properties of proteins, permit efficient algorithms, and produce accurate predictions. Moreover, we have shown here that the Bayesian framework is naturally generalized to model non-local interactions in protein folding. As an example, we have presented a simple model for β-strand pairing, and a Markov chain Monte Carlo algorithm for inference, and have demonstrated this approach on example sequences. Further work on modeling and evaluation for this problem is underway (Schmidler, 2000). The ability to predict tertiary contacts between β-sheets represents a potentially important step in going beyond traditional secondary structure prediction towards the goal of full 3D structure prediction.

Acknowledgments

SCS was partially supported by NLM training grant LM-07033 and NCHGR training grant HG-00044-04 during portions of this work. JSL is partially supported by NSF grants DMS-9803649 and DMS-0094613. DLB is supported by NHGRI grant HGF02235-07.

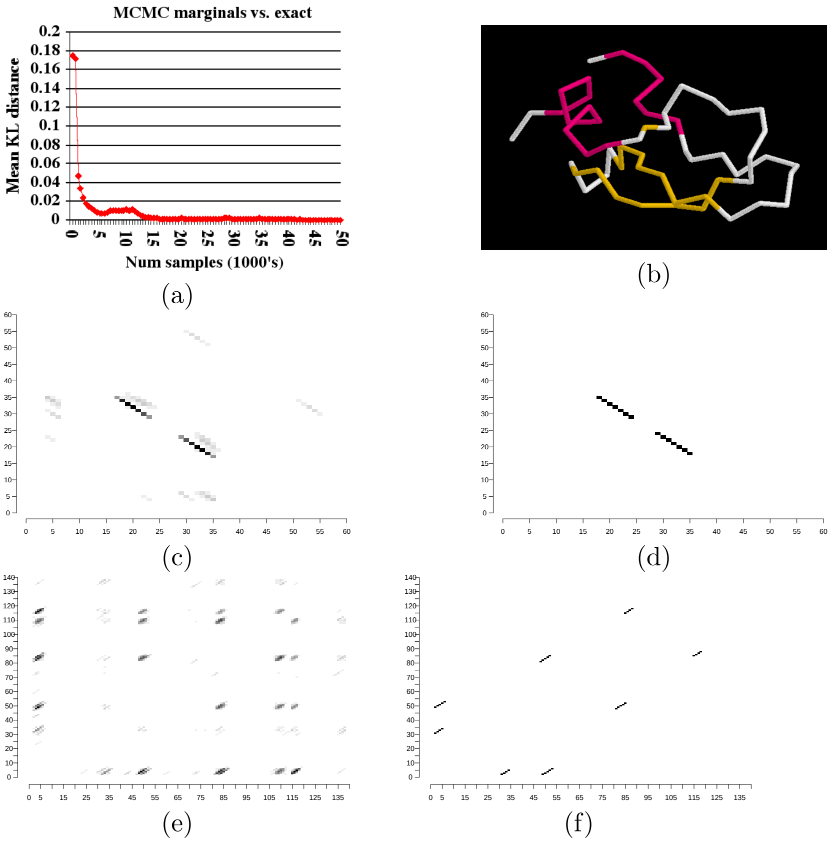


FIGURE 5. (a) Convergence of MCMC simulation to exact calculations. Plot is mean Kullback-Leibler (KL) divergence between marginal distributions $P(T_{R_{[i]}} \mid R, \theta)$ obtained from exact and MCMC calculations for a protein sequence, against number iterations (each iteration 1 full scan). KL divergence between two probability distributions $\mathbf{p}$ and $\mathbf{q}$ is defined as $KL(\mathbf{p}, \mathbf{q}) = \sum_i p_i \log(\frac{p_i}{q_i})$. (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.

Bibliography

Asai, K., Hayamizu, S., and Handa, K. (1993). Prediction of protein secondary structure by the hidden Markov model. *Comp. Appl. Biosci.*, 9(2):141–146.

Aurora, R. and Rose, G. D. (1998). Helix capping. *Prot. Sci.*, 7:21–38.

Baldwin, R. L. and Rose, G. D. (1999). Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.*, 24:26–33.

Barton, G. J. (1995). Protein secondary structure prediction. *Curr. Opin. Struct. Biol.*, 5:372–376.

Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W., and Swaminathan, S. (1999). Structural genomics: Beyond the Human Genome Project. *Nat. Genet.*, 23:151–157.

Cohen, B. I., Presnell, S. R., and Cohen, F. E. (1993). Origins of structural diversity within sequentially identical hexapeptides. *Prot. Sci.*, 2:2134–2145.

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. (1998). New goals for the U.S. Human Genome Project: 1998-2003. *Science*, 282:682–689.

Dill, K. A. (1999). Polymer principles and protein folding. *Prot. Sci.*, 8:1166–1180.

Eyrich, V. A., Standley, D. M., and Friesner, R. A. (1999). Prediction of protein tertiary structure to low resolution: Performance for a large and structurally diverse test set. *J Mol. Biol.*, 288:725–742.

Fischer, D. and Eisenberg, D. (1996). Protein fold recognition using sequence-derived predictions. *Prot. Sci.*, 5:947–955.

Frishman, D. and Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Prot. Eng.*, 9(2):133–142.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1996). *Markov Chain Monte Carlo in Practice.* Chapman & Hall.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–32.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Hubbard, T. J. and Park, J. (1995). Fold recognition and ab initio structure predictions using hidden Markov models and β-strand pair potentials. *Proteins: Struct. Funct. Genet.*, 23:398–402.

Hutchinson, E. G., Sessions, R. B., Thornton, J. M., and Woolfson, D. N. (1998). Determinants of strand register in antiparallel β-sheets of proteins. *Prot. Sci.*, 7:2287–2300.

Kabsch, W. and Sander, C. (1984). On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. *Proc. Natl. Acad. Sci. USA*, 81(4):1075–1078.

King, R. D. and Sternberg, M. J. E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Prot. Sci.*, 5:2298–2310.

Klingler, T. M. and Brutlag, D. L. (1994). Discovering structural correlations in α-helices. *Prot. Sci.*, 3:1847–1857.

Krogh, A. and Riis, S. K. (1996). Prediction of beta sheets in proteins. In Touretzky DS, Mozer MC, H. M., editor, *Advances in Neural Information Processing Systems 8*. MIT Press.

Lifson, S. and Sander, C. (1980). Specific recognition in the tertiary structure of β-sheets of proteins. *J Mol. Biol.*, 139:627–639.

Liu, J. S. (1996). Peskun's theorem and a modified discrete-state Gibbs sampler. *Biometrika*, 83:681–682.

Minor, D. L. J. and Kim, P. S. (1996). Context-dependent secondary structure formation of a designed protein sequence. *Nature*, 380:730–734.

Monge, A., Friesner, R. A., and Honig, B. (1994). An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc. Natl. Acad. Sci. USA*, 91:5027–5029.

Montelione, G. T. and Anderson, S. (1999). Structural genomics: Keystone for a Human Proteome Project. *Nat. Struct. Biol.*, 6:11–12.

Neumaier, A. (1997). Molecular modeling of proteins and mathematical prediction of protein structure. *SIAM Rev.*, 39(3):407–460.

Russell, R. B., Copley, R. R., and Barton, G. J. (1996). Protein fold recognition by mapping predicted secondary structures. *J Mol. Biol.*, 259:349–365.

Schmidler, S. C. (2000). *Statistical Models and Monte Carlo Methods for Protein Structure Prediction.* PhD thesis, Stanford University.

Schmidler, S. C., Liu, J. S., and Brutlag, D. L. (2000). Bayesian segmentation of protein secondary structure. *J. Comp. Biol.*, 7(1):233–248.

Stultz, C. M., White, J. V., and Smith, T. F. (1993). Structural analysis based on state-space modeling. *Prot. Sci.*, 2:305–314.