Stochastic Segment Interaction Models for Biological Sequence Analysis

Scott C. Schmidler*, Jun S. Liu†, Douglas L. Brutlag‡

Abstract

We introduce a class of probability models for sequences of random variables with complex long-range dependency structure, called *stochastic segment interaction models*, motivated by problems arising in the analysis of biopolymer sequence data. We generalize and extend previous work in this area, and make explicit the relations to existing literature on hidden Markov models (HMMs) and "generalized" HMMs. We show that this class of models allows for incorporation of non-local interaction information in biological sequence analysis. We demonstrate this approach by developing models for prediction of 3D contacts in protein sequences using models for amino acid dependencies in β-sheets. We provide algorithms for Bayesian inference on these models via dynamic programming and Markov chain Monte Carlo simulation. Results are presented from an application to protein structure prediction from sequence.

Keywords: Hidden Markov models, segmentation, Bayesian methods, protein structure prediction, β-sheets, sequence analysis

1 Introduction

Sequential data exhibiting long-range dependencies arises in many contexts, including time series analysis, longitudinal data analysis, signal processing and speech recognition, and biological sequence analysis. Such problems are intrinsically challenging due to the inherent difficulties in defining, estimating, validating, and computing with models for joint distributions over large numbers of random variables. A key aspect of analyzing long-range dependency is the development of statistical models with limited, structured dependency

*Scott C. Schmidler is Assistant Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251. Tel: (919) 684-8064; Fax: (919) 684-8594; Email: schmidler@stat.duke.edu

†Jun S. Liu is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138. Ph: (617) 495-1600; Fax: 496-8057; Email: jliu@stat.harvard.edu

‡Douglas L. Brutlag is Professor, Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305-5307. Tel: (650) 723-6593; Fax (650) 723-6783; Email: brutlag@stanford.edu.

for distant observations. Much effort has been devoted to continuous stochastic processes and time series analysis, but may be less suitable for problems where the observed long-range dependencies arise from other physical processes. Such is the case in analysis of biological sequence data (e.g. DNA, RNA, and proteins) where the random variables are categorical and the long-range dependencies arise via physical interactions in three-dimensional space.

In this paper we introduce a class of probability models for sequences of random variables with complex long-range dependency structure, which we call *stochastic segment interaction models*. These models are especially motivated by problems arising in the analysis of biopolymer sequence data. We review and extend previous work on *stochastic segment models* (Schmidler et al., 2000; Ostendorf et al., 1996; Kulp et al., 1996; Burge and Karlin, 1997), and formalize the relations between them. We then introduce stochastic segment interaction models, a generalization of stochastic segment models which allows for incorporation of long-range dependency structure. Bayesian analysis via prior distributions on these models is discussed, and algorithms for computing with these models are provided using dynamic programming recursions and Markov chain Monte Carlo (MCMC) simulation. Applications to modeling of protein and RNA sequences are described, and empirical results are reported for a difficult problem in the prediction of protein structure from sequence.

2 Non-local interactions in biopolymer sequences

Linear polymers are large molecules made up of strings of component submolecules arranged into a chain. Among the best known biological polymers are DNA, RNA, and proteins, which play critical roles in sustaining and reproducing life in biological organisms. Because they are linear (unbranched), they are commonly represented by a linear sequence of letters, in which each letter represents the identity of the component at that position. For example, a DNA sequence represents successive nucleotides:

```
ACCGTACATCGAGAAGTCCTAGATTATACTA
```

while a protein sequence represents successive amino acids:

```
DGVAEITIKLPRHRNALSVKAMQEVTDALNRAEEDDSVGAVMITGAE        (1)
```

Much has been learned from computational and statistical analysis of such character strings. However this representation disguises the fact that these biopolymers are large, complex molecules which adopt complicated geometric shapes in order to perform their many roles (see Figure 1a). In understanding the function and evolution of these molecules and their roles in biological processes, including topics such as protein folding, complex formation, ligand binding, chromatin organization, structure, and molecular design, it is often necessary to understand aspects of this molecular structure. An important property of such sequences is the formation of *non-local* dependencies, via physical interactions between positions far apart in the linear sequence but close in 3-dimensional space.

Non-local interactions in protein structure Proteins fold into complex structures which can be decomposed into elements of two major secondary structure classes known as α-helices and β-sheets (Schmidler et al., 2001). β-sheets are formed by hydrogen bonding between distinct regions of the polymer separated arbitrarily far in the linear sequence, as shown in Figure 1c. Many other such non-local interactions occur in proteins, including helical bundles (Figure 1b) and coiled coils, hydrophobic contacts, salt bridges, and disulfide bonds.

Non-local interactions in RNA structure Another well-known example of long-range interactions is the formation of secondary structure in RNA. The most recognizable elements of RNA secondary structure are the intramolecular double helices. These helices are similar to the well known double-helix of DNA, but involve base-pairing between distant regions of the same sequence rather than between distinct sequences. RNA secondary structure is perhaps the best-studied example of non-local interactions in biopolymers, and can be described using stochastic grammars (see Section 4.3). An example RNA secondary structure is shown in Figure 1c.

3 Stochastic segment models

3.1 Notation

Let $X = (X_1, \ldots, X_n)$ be an observed sequence of random variables taking values in a finite alphabet $X_i \in \mathcal{A}_X$, and let $Y = (Y_1, \ldots, Y_n)$ be an sequence of unobserved states for each X_i, taking values in finite alphabet $Y_i \in \mathcal{A}_Y$ (see Figure 2). We are particularly interested in the case where X is the sequence of a biological polymer (protein or nucleic acid), and $\mathcal{A}_X$ is the set of 20 naturally-occurring amino acids or 4

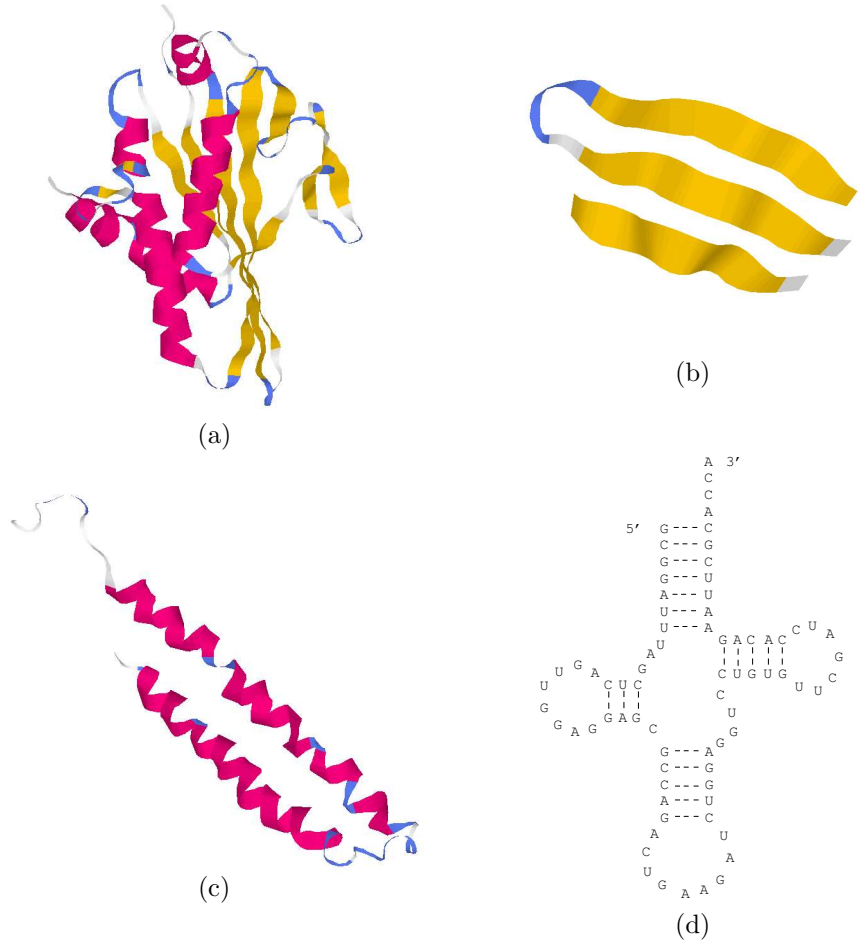


Figure 1: Non-local interactions in biopolymer sequences. (a) A protein (b) A β-sheet (D sheet from mouse immunoglobulin heavy chain (1a6w)) (c) An α-helical hairpin (residues 1-95 from a phenylalanyl-tRNA synthetase (1eiy)) (d) an RNA secondary structure: yeast Phe tRNA (1ehz)

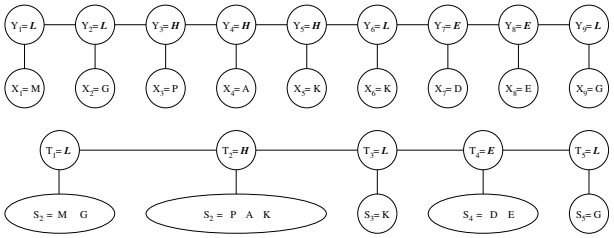


Figure 2: (a) HMM representation of a sequence of observed random variables and corresponding unobserved states. (b) Segment representation of unobserved states.

nucleotide bases. The alphabet $\mathcal{A}_Y$ is application dependent, but possibilities include backbone conformation in proteins (Asai et al., 1993; Stultz et al., 1993; Schmidler et al., 2000) or genome structure in DNA (Churchill, 1989; Stormo and Haussler, 1994; Kulp et al., 1996; Burge and Karlin, 1997; Braun and Muller, 1998).

Segmentations: An alternative representation for the unobserved state sequence Y is as a sequence of *segments*, defined as $(value, length)$ pairs obtained by grouping consecutive Y_i's with identical values. We denote this sequence by $\mathcal{S} = (S_1, \ldots, S_m) = ((T_1, \ell_1), \ldots, (T_m, \ell_m))$, and refer to $\mathcal{S}$ as a *segmentation* of the sequence X.

Although $\mathcal{S}$ fully specifies a unique segmentation, it is convenient to introduce the following (slightly redundant) additional notation:

(i) $m = |\mathcal{S}|$, the number of segments

(ii) $s_i = 1 + \sum_{j<i} \ell_j$, the first sequence position of segment i

(iii) $e_i = s_i + \ell_i - 1$, the last sequence position of segment i

An example of this notation is given in Figure 2. The segment locations $\{s_i\}_{i=1}^m$ (or equivalently $\{e_i\}_{i=1}^m$) are *changepoints* in statistical terms (Barry and Hartigan, 1993; Stephens, 1994; Braun and Muller, 1998). Note the implicit constraints $s_1 = 1$, $e_m = n$, and $s_i = e_{i-1} + 1$ for $i = 2, \ldots, m$.

Segment interactions: (Schmidler, 2002) introduced the notion of *segment interactions*. A segment interaction specifies a relation between two or more segments in a segmentation.

In general terms, given a sequence of random variables X and a segmentation $\mathcal{S}$, we define a *segment interaction* I to be a pair $(\mathcal{H}, \eta)$ where $\mathcal{H} = (H_1, \ldots, H_k)$ indexes a subset of segments and η is a set of parameters specifying the precise pattern of interaction. (In the most general form, $\eta = \{h_i\}_{i=1}^{2^k}$ where h_i are

parameters specifying the interactions of the i^{th} subset of $\mathcal{H}$.) There may be multiple segment interactions for a sequence, and we denote the set of interactions as $\mathcal{I} = \{I_i\}_{i=1}^{p}$. An interaction I is defined to be maximal (the maximal clique in a triangulated graphical model defined on segments of the sequence, see Figure 3), so $\mathcal{H}_i \cap \mathcal{H}_j = \emptyset$ holds $\forall i \neq j$. If we define any non-interacting segment in $S_i \in \mathcal{S}$ to be an interaction of size 1 (so $I_j = (\{i\}, \emptyset)$ and $k_j = |\mathcal{H}_j| = 1$), then $\mathcal{I}$ induces a partition of $\mathcal{S}$ of size p, yielding $1 \leq p \leq m$ mutually exclusive and exhaustive subsets with $\mathcal{S} = \bigcup_{i=1}^{p} \mathcal{H}_i$. We refer to the set $(\mathcal{S}, \mathcal{I})$ as an *interacting segmentation*.

As a concrete example of a segment interaction (see Section 9.2), $\mathcal{I}$ may denote the set of β-sheets in a protein, $\mathcal{H}_i$ the set of β-strands making up the i^{th} sheet, and η_i parameters specifying the relative orientation and register of neighboring strands. Alternatively, $I \in \mathcal{I}$ might represent a helical bundle or other super-secondary structure, or a helix or pseudo-knot in RNA structure.

3.2 Stochastic segment models

In previous work we have developed a class of probability models defined on segmentations of the form (Schmidler et al., 2000):

$$P(X, \mathcal{S} \mid \theta, \gamma) \propto P(\mathcal{S} \mid \gamma) \prod_{j=1}^{m} P(X_{[s_j:e_j]} \mid \mathcal{S}, \theta) \tag{2}$$

The key assumption of (2) is the conditional independence of positions X_i occurring in different segments, given a segmentation $\mathcal{S}$. (Note that *marginally*, the observed sequence X has a complex dependency structure.) The segment likelihoods $P(X_{[s_i:e_i]} \mid \mathcal{S}, \theta)$ capture *local* dependencies through the joint distribution of intra-segment positions, and may be of general form. These models have been shown to be particularly appropriate for modeling aspects of protein secondary structure and hence for Bayesian protein structure prediction (Schmidler et al., 2000). When the segmentation prior $P(\mathcal{S} \mid \gamma)$ in (2) is factored appropriately, posterior quantities are obtained exactly by efficient algorithms (see Section 7).

A slightly less general form of (2) is discussed in (Ostendorf et al., 1996) under the name of *stochastic segment models*, and we adopt this terminology here. The relations between these and other models developed in the speech recognition and bioinformatics communities are discussed in Section 4.

For many applications such as those discussed in Section 2, parameters (θ, γ) may be estimated from data $X^1, \ldots X^r$ where $\mathcal{S}^i$ is observed for each X^i. Then the log-likelihood

$$L\left(X^1, \ldots, X^r; \theta, \gamma\right) = \sum_{i=1}^{r} \log P(\mathcal{S}^i \mid \gamma) + \sum_{j=1}^{m} \log P(X^i_{[s_j:e_j]} \mid \mathcal{S}, \theta) \tag{3}$$

decomposes according to the segment conditional independence, and parameters (θ, γ) may be estimated for each segment type separately using MLE or Bayesian approaches.

In the absence of fully-observable $\mathcal{S}^i$'s (3) may be viewed as the complete-data log-likelihood, and standard missing-data estimation methods such as EM (Depmster et al., 1977) or MCMC (Gilks et al., 1996) may be applied. Thus in what follows, we focus instead on recovering the segmentation $\mathcal{S}$ itself given (θ, γ). We formulate this as a Bayesian inference problem:

$$P(\mathcal{S} \mid X, \theta, \gamma) = \frac{P(\mathcal{S} \mid \gamma) \prod_{j=1}^{m_{\mathcal{S}}} P(X_{[s_j:e_j]} \mid \mathcal{S}, \theta)}{\sum_{\mathcal{S}'} P(\mathcal{S}' \mid \gamma) \prod_{j=1}^{m_{\mathcal{S}'}} P(X_{[s_j:e_j]} \mid \mathcal{S}', \theta)}$$

and either condition on or integrate out the parameters (θ, γ). In what follows we suppress the dependence on these parameters.

3.3 Stochastic segment interaction models

The class of stochastic segment models described by (2) can be generalized to a much larger class of models involving the segment interactions introduced in Section 3.1. We write a joint distribution over the set of *interacting segmentations* in the following form:

$$P(X, \mathcal{S}, \mathcal{I}) \propto P(\mathcal{S}, \mathcal{I}) \prod_{i=1}^{p} P(\{X_{[s_j:e_j]}\}_{S_j \in \mathcal{H}_i} \mid \mathcal{S}, \mathcal{I}) \tag{4}$$

which factors by conditional independence of *sets of interacting* segments. We refer to this new class of models, first introduced in (Schmidler, 2002), as *stochastic segment interaction models* (SSIMs). Here models of inter-segment sequence dependency may be introduced by inclusion of *joint-segment likelihoods*, replacing the terms

$$P(X_{[s_j:e_j]} \mid S_j) \quad \text{and} \quad P(X_{[s_k:e_k]} \mid S_k)$$

for two interacting segments S_j and S_k in the product of (2) above with a joint term:

$$P(X_{[s_j:e_j]}, X_{[s_k:e_k]} \mid S_j, S_k) \tag{5}$$

Hence positions X_i in different segments may be made conditionally *dependent* by introducing an interaction I containing the two segments $(S_j, S_k \in H_I)$. Thus arbitrary joint-segment distributions for segment pairs

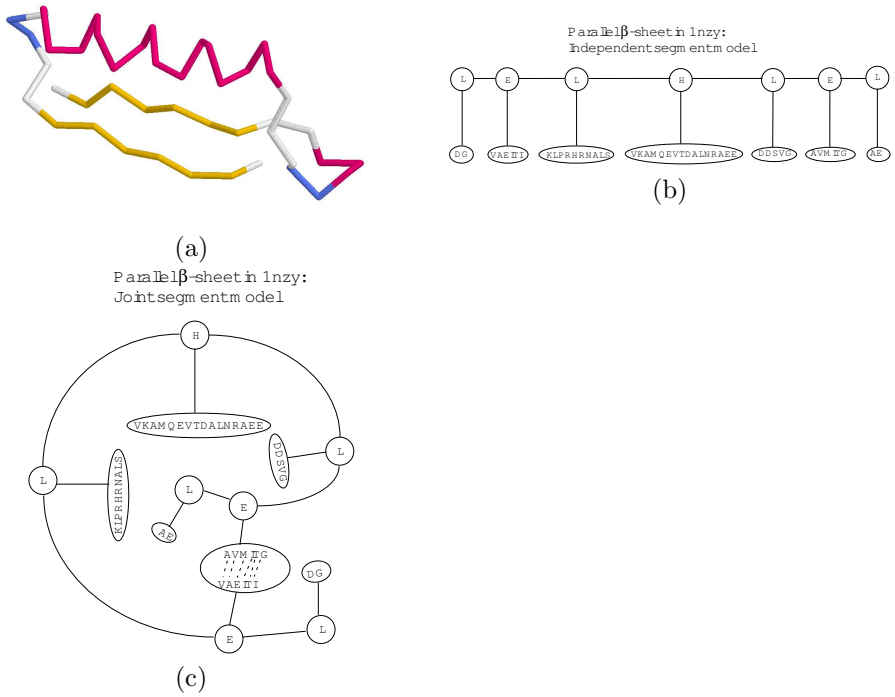


Figure 3: Segment representation of a sequence including segment interactions. (a) A parallel β-sheet from (1nzy_A) (b) Graphical representation of SSM model of this sequence. (c) Graphical representation of SSIM model of this sequence. Here $\mathcal{I} = \{I_1\}$ and $I_1 = (\{2, 5\}, \eta_1)$.

may be incorporated into the model. The extension to joint distributions on three or more segments (as may be required for 4-helix bundles or β-sheets, for example) is obvious. Figure 3 shows an example of such a joint distribution. This class of models is sufficiently general to capture the significant non-local dependencies in protein sequences; see Sections 4.3 and 9.2 for examples. When $p = m$, (4) reduces to (2), so this class of models strictly generalizes those developed previously.

As will be described in Section 7, many models of the form (2) enjoy nice computational properties. In contrast, models of the form (4) typically present significant computational difficulties. While the joint distribution (4) is easily evaluated for any fixed segmentation $\mathcal{S}$ of X, calculation of relevant predictive quantities under this model rarely permits efficient exact algorithms. Approximation algorithms based on Monte Carlo simulation are developed in Section 8.

4 Hidden Markov models and stochastic segment models

The class of models described by (2) above has close ties to other stochastic sequence models, and it is helpful to make these explicit.

4.1 Hidden Markov models

Hidden Markov models (HMMs) have been widely used in bioinformatics (Churchill, 1989; Baldi et al., 1994; Krogh et al., 1994; Asai et al., 1993; Stultz et al., 1993; Eddy, 1996), as well as many other areas of engineering and statistics (Rabiner, 1989; MacDonald and Zucchini, 1997).

Letting Y_i again be the hidden state at position i a HMM may be written in the form:

$$P(X, Y) = \prod_{i=1}^{n} P(Y_i \mid Y_{i-1}) P(X_i \mid Y_i) \tag{6}$$

where throughout we will let $P(Y_1 \mid Y_0)$ denote $P_0(Y_1)$, the initial distribution of the hidden Markov chain. Rearranging (6) into segment form, we obtain:

$$P(X, \mathcal{S}) = \prod_{i=1}^{m} P(Y_i \mid Y_{i-1}) P(Y_i \mid Y_i)^{\ell_i - 1} \prod_{j=s_i}^{e_i} P(X_j \mid Y_i) \tag{7}$$

From (7) we observe that HMMs are a special case of stochastic segment models with two additional strong assumptions imposed:

(i) *Length distributions are geometric:* Segment lengths follow a geometric distribution with parameter $p_T = P(Y_i = T \mid Y_{i-1} = T)$, so

$$P(\ell_i = k \mid T_i) \propto p_{T_i}^k \tag{8}$$

(ii) *Positions are conditionally iid:* All observed sequence positions are conditionally independent given the segments, even those in the same segment:

$$P(X_i | \mathcal{S}, X_{j \neq i}) = P(X_i | \mathcal{S}) \tag{9}$$

and positions within a given segment are identically distributed:

$$P(X_{[s_i, e_i]} \mid S_i) = \prod_{j=s_i}^{e_i} P(X_j \mid T_i) \tag{10}$$

4.2 Hidden semi-Markov models, generalized HMMs, and stochastic segment models

Violations of assumptions (8-10) in applications such as bionformatics and speech recognition have given rise to generalizations of HMMs. These include incorporation of intra-segment position dependence and arbitrary, type-dependent segment length distributions. These models are of the form:

$$P(X, \mathcal{S}) \propto \prod_{j=1}^{m} P(X_{[s_j:e_j]} \mid T_j) P(T_j \mid T_{j-1}) P(\ell_j \mid T_j) \tag{11}$$

and appear under various names, including *generalized* HMMs (Stormo and Haussler, 1994; Kulp et al., 1996; Burge and Karlin, 1997) and *stochastic segment* (Ostendorf et al., 1996) or *Bayesian segmentation* (Schmidler et al., 2000) models. Imposing (9) and (10) but relaxing (8) gives the special case of "explicit state duration density" HMMs or hidden *semi*-Markov models (HSMM) (Russell and Moore, 1985; Levinson, 1986; Rabiner, 1989).

All of these models differ from HMMs by changing the (prior) distribution of $\mathcal{S}$ from a Markov process

$$P(\mathcal{S}) = \prod_{j=1}^{n} P(Y_j \mid Y_{j-1}) \tag{12}$$

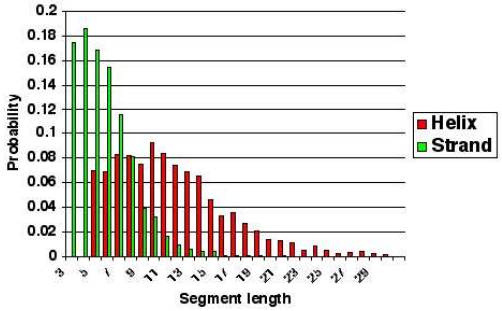
to a *semi*-Markov process:

$$P(\mathcal{S}) = \prod_{j=1}^{m} P(T_j \mid T_{j-1}) P(\ell_j \mid T_j) \tag{13}$$

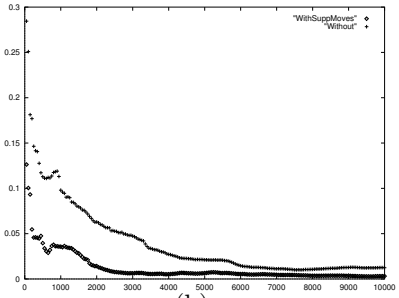
In this paper we adopt the term *stochastic segment models* (SSMs) to denote the slightly more general class of models described by (2) which allow general priors $P(\mathcal{S})$. We use *generalized HMMs* (GHMMs) to refer to the special case where the prior distribution $P(\mathcal{S})$ is of the form (11).

Explicit modeling of segment length has proven useful in bioinformatics for accounting for differences in intron/exon length in eukaryotic DNA (Kulp et al., 1996; Burge and Karlin, 1997) and different types of secondary structure in proteins (Schmidler et al., 2000). Figure 4a shows the relative length frequencies of two types of protein secondary structure segments in the dataset described in Section 9.

The advantage of GHMMs and SSMs over HSMMs lies in the ability to model conditional dependencies between intra-segment sequence positions. Important dependencies exist between positions within secondary structure elements of proteins and within coding and regulatory regions of DNA. Standard methods for protein structure prediction and gene recognition incorporate these dependencies to improve predictive



(a)



(b)

Figure 4: (a) Histogram of observed segment lengths for protein α-helices (red) and β-strands (green). (b) MCMC convergence for an example sequence (5nul) with and without supplemental moves. Shown is mean Kullback-Leibler divergence between marginals $P(T_{X_{[i]}} \mid \theta)$ obtained by exact and MCMC calculations for SSM.

performance (see above references for details). Generally speaking, we may write the joint distribution of intra-segment positions in log-linear form as a Gibbs random field potential:

$$\log P(X_{[s:e]} \mid \cdot) \propto$$

$$\sum_{s \leq i \leq e} f_i(X_i) + \sum_{s \leq (i,j) \leq e} g_{i,j}(X_i, X_j) + \sum_{s \leq (i,j,k) \leq e} h_{i,j,k}(X_i, X_j, X_k) + \ldots \quad (14)$$

from which it is clear that HMM and HSMM models, which set all interaction coefficients to zero, are quite restrictive.

4.3 Stochastic segment interaction models

As described in Section 3.1, SSIMs go further by relaxing the conditional independence assumption for *inter*-segment sequence positions using a structured notion of segment interaction. In practice, the overly general form of SSIMs given by (4) must be traded off against model complexity by modeling strong dependencies and ignoring others. To see this generality, we note that random fields defined by pair potentials arise as special cases of SSIMs (by taking $m = n$ and $p = 1$). Pair potentials are common in physical models, and empirical pair potentials have been used for the problem of β-sheet prediction in proteins described in Section 9.2 (Hubbard, 1994).

4.3.1 RNA folding and stochastic grammars

A class of models developed in the computer science and bioinformatics literature known as *stochastic context-free grammars* (SCFGs) also fall into the SSIM framework. SCFGs have been successfully applied to RNA secondary structure prediction and speech recognition (Lari and Young, 1991; Sakakibara et al., 1994; Eddy and Durbin, 1994). We give a brief overview overview of SCFGs here; a detailed treatment in the context of RNA structure prediction is given in (Durbin et al., 1998). SCFGs also encompass many statistical mechanical models of RNA secondary structure parameterized using experimentally determined energies (Zuker and Stiegler, 1981; Zuker and Sankoff, 1984; Zuker, 1989; McCaskill, 1990).

The secondary structure of an RNA sequence can be represented by a set of ordered base pairs $H = \{(i, j)\}$ under the constraints that for any two pairs $(i_1, j_1), (i_2, j_2) \in H$ such that $i_1 \leq i_2$, we have:

$$i_1 = i_2 \Leftrightarrow j_1 = j_2 \tag{15}$$

$$i_2 < j_1 \Rightarrow i_1 < i_2 < j_2 < j_1 \tag{16}$$

Biologically these constraints say that each nucleotide may participate in only one base pair, and restrict the structure from forming pseudo-knots. Mathematically, they say that the structure may be drawn as a planar graph, and may be represented by a context-free grammar (Searls, 1993).

Given a secondary structure H of this form, the probability of a ribonucleotide sequence X may be written as a Gibbs random field (Kindermann and Snell, 1980):

$$P(X \mid H) \propto \exp(-U(X, H)/kT) \tag{17}$$

with

$$U(X, H) = \sum_{i \notin H} f(X_i) + \sum_{(i,j) \in H} g(X_i, X_j)$$

a pair potential where $g(X_i, X_j)$ is the free energy obtained by pairing nucleotides i and j. Typically U is defined over neighboring pairs to account for effects such as base stacking.

An useful representation for models of the form (15,17) is as a stochastic context-free grammar. Secondary structures obeying the constraints (15) can be generated by a context-free grammar, defined by a set of *rewrite*

rules of the form:

$$A \to Aa \mid Au \mid Ac \mid Ag \mid B \mid \epsilon \qquad \text{and} \qquad B \to aAu \mid gAc$$

where A, B are *non-terminals* which are instantiated in various ways by applying the replacement on the right hand side of the rules. The nucleotides a,u,c,g and null string ϵ are *terminals*, and repeated application of these substitution rules generates sequences of terminals:

$$A \to B \to aAu \to aBu \to aaAuu \to aaBuu \to aacAguu \to aacaguu$$

in this case generating a hairpin loop with three base pairs formed between aac and guu. A *stochastic* CFG is obtained by assigning probabilities to the various replacement rules, making the non-terminals into random variables, and resulting in a stochastic generative model for strings. HMMs can also be written in this form and result in a special case of SCFGs known as stochastic *regular* grammars; other types of rewrite rules give rise to other classes of grammars see (Hopcroft and Ullman, 1979) for details. Formal grammars are a standard tool of linguistics and theoretical computer science.

The set of rewrite rule applications by which an observed sequence is generated is known as a *parse tree*; the parse tree is analogous to the hidden state sequence in an HMM. As with HMMs, efficient algorithms exist for computing with SCFGs; for example, the marginal likelihood of an observed sequence may be calculated by summing over all possible parse trees in $O(n^3)$ steps using the *inside-outside* algorithm (Lari and Young, 1990), a generalization of the forward-backward algorithm for HMMs (Rabiner, 1989).

For current purposes, it is sufficient to note that the conditional probability of a sequence given a parse tree may be written in the form:

$$\begin{aligned} P(aacuu \mid \mathcal{T}) &= P(A \to B)P(B \to aAu)P(B \to aAu)P(A \to Ac)P(A \to \epsilon) \\ &= P(au \mid \mathcal{T})P(au \mid \mathcal{T})P(c \mid \mathcal{T}) \end{aligned} \tag{18}$$

and that rewrite rules of the form $A \to aBb$ lead to stochastic dependencies between distant positions in the sequence. However, these long-range dependencies are restricted to those generated by the parse tree of a CFG. Thus correlated positions must obey constraints of the form (15).

Models of form (17,18) are special cases of SSIMs (4) which use only a restricted subset of non-local interactions. The primary advantage of such restrictions is the amenability to efficient calculation via inside-

outside algorithms. The tendency of many molecular interactions such as protein β-sheets to violate the constraints (15) was a major motivation in the the development of SSIMs. SSIMs may also have applications to RNA folding involving pseudo-knots, but we do not pursue this here.

Hierarchy of models: In general, we may view these models as a hierarchy of model complexity, each subsuming the previous:

$$HMM \subset HSMM \subset SSM \subset SCFG \subset SSIM$$

Choosing the appropriate class of models and testing for fit falls into the domain of Bayesian model selection, and will be described elsewhere.

5 Priors on segmentations

To complete the SSM model we must specify the prior probability distribution on segmentations $P(\mathcal{S})$. As mentioned in Section 4, our definition of SSMs (2) generalizes GHMMs (11) by allowing alternative forms for $P(\mathcal{S})$. From a Bayesian perspective, GHMMs (and the SSMs of Ostendorf et al., 1996) are SSMs with a specific form of *segmentation prior*.

It is often convenient to specify a segmentation prior in the form:

$$P(\mathcal{S}) = P(m)P(S_1, \ldots, S_m \mid m) \tag{19}$$

where m is the number of segments. As seen in Section 4.2, GHMMs (and therefore HSMMs and standard HMMs) implicitly assume a prior of this form, with $P(S_1, \ldots, S_m \mid m)$ factored as a (semi-)Markov process. This embodies certain assumptions on (19):

(i) *Uniform number segments:* We have

$$P(m) \propto 1$$

so $P(\mathcal{S})$ is improper (m is unbounded), but a proper posterior is obtained by conditioning on an observed sequence X of finite length n.

(ii) *Markovian segment types:* The sequence of segment types is given a Markov or nearest-neighbor dependency structure

$$P(T_1, \ldots, T_m \mid m) = \prod_{j=1}^{m} P(T_j \mid T_{j-1}) \tag{20}$$

(iii) *Lengths iid:* The length distributions of all segments are conditionally independent, and are identically distributed for all segments of the same type:

$$P(\ell_1, \ldots, \ell_m \mid m, T_1, \ldots, T_m) = \prod_{j=1}^{m} P(\ell_j \mid T_j) \tag{21}$$

The wide application of HMMs in practice suggests that these assumptions are generally reasonable; however, they often go unnoticed and may be inappropriate in some domains. Below we describe several other possible forms of segmentation priors.

5.1 Alternative segmentation priors

SSMs are not limited to the priors adopted implicitly in GHMMs. We briefly outline some natural alternatives.

Uniform prior: A simple approach considers a prior that is uniform over segmentations:

$$P(\mathcal{S}) \propto 1 \tag{22}$$

yielding the joint distribution

$$P(X, \mathcal{S}) \propto \prod_{j=1}^{m} P(X_{[s_j:e_j]} \mid \mathcal{S}) \tag{23}$$

Again (22) is improper but yields a proper posterior given finite n. Prior (22) is adopted implicitly in related work on DNA and protein sequence segmentation (Auger and Lawrence, 1989; Stormo and Haussler, 1994; Liu and Lawrence, 1996). The resulting MAP segmentation (see Section 6) may be interpreted as a maximum likelihood segmentation of the sequence X.

Semi-Markov process prior: The semi-Markov process prior (13) described above has been used by (Snyder and Stormo, 1993; Kulp et al., 1996; Burge and Karlin, 1997) for DNA and (Schmidler et al., 2000) for proteins. It may also be combined with any proper marginal prior $P(m)$, yielding the joint distribution:

$$P(X, \mathcal{S}) = P(m) \prod_{j=1}^{m} P(X_{[s_j:e_j]} \mid \mathcal{S}) P(T_j \mid T_{j-1}) P(\ell_j \mid T_j) \tag{24}$$

Choice of $P(m)$ may affect computational complexity of inference; see Section 7.

Sequence-specific prior: Priors (22) and (13) provide a generative model for segmentations. Alternatively we may consider *sequence-specific* priors over segmentations of a particular sequence X. This fixes the

sequence length n as part of the model rather than an observed quantity, assigning prior mass only to segmentations satisfying the constraint $\sum_{j=1}^{m} \ell_j = n$. This approach is used by (Liu and Lawrence, 1999), who adopt a uniform prior on m-segmentations conditional on m. In our notation this becomes:

$$P(\mathcal{S} \mid n) \propto P(m) \binom{n-1}{m-1}^{-1} (|\mathcal{A}_T| - 1)^{-(m-1)} / |\mathcal{A}_T| \tag{25}$$

Similarly, one can adapt the uniform prior (22) to the sequence-dependent case, where the only effect of conditioning on n:

$$P(\mathcal{S} \mid n) \propto 1 \tag{26}$$

is to make the prior proper.

Other segment-decomposable priors: The above priors all share an important property, which we call *segment decomposability*:

Definition. A probability distribution P on $\mathcal{S}$ is *segment-decomposable* if

$$P(\mathcal{S}) \propto \prod_{j=1}^{m} f(e_{j-1}, e_j, T_{j-1}, T_j)$$

for some f independent of j.

Note that to satisfy this definition, the distribution $P(m)$ must factor accordingly (e.g. geometric or uniform). When it does not, we may obtain a similar property by conditioning on m:

Definition. P is *conditionally segment-decomposable given* Z if

$$P(\mathcal{S} \mid Z) \propto g(Z) \prod_{j=1}^{m} f(e_{j-1}, e_j, T_{j-1}, T_j \mid Z)$$

For SSMs (2) the posterior $P(\mathcal{S} \mid X)$ is (conditionally) segment decomposable *if and only if* the prior $P(\mathcal{S})$ is, a sort of conjugacy property. Segment decomposability of the posterior dramatically affects computational complexity of inference (see Section 7). All priors given above are examples of (conditionally) segment-decomposable priors.

General segmentation priors: It is easy to construct desirable priors which are not segment-decomposable. For example, one may wish to specify a prior on the percentage of various types of segments, such as the

percentage secondary structure content of proteins:

$$P(\mathcal{S}) = P(\%H, \%E, \%L \in T \mid m)P(m) \tag{27}$$

Priors involving such global features of a segmentation impose a non-decomposable structure on the joint distribution (2) which complicates posterior computation (see Section 8).

5.2 Priors on segment interactions

In specifying SSIM models, we require priors involving interactions of the form $P(\mathcal{S}, \mathcal{I})$. Again, we distinguish several types:

Uniform: Analogous to (22), we may adopt a uniform prior on interacting segmentations:

$$P(\mathcal{S}, \mathcal{I}) \propto 1 \tag{28}$$

which yields a maximum likelihood interacting segmentation.

Conditionally uniform: Any of the priors discussed in Section 5.1 may be extended by a conditionally uniform prior on segment interactions, yielding:

$$P(\mathcal{S}, \mathcal{I}) \propto P(\mathcal{S})/c(\mathcal{S}) \tag{29}$$

where $c(\mathcal{S})$ is the number of possible interactions formed on segments in $\mathcal{S}$, and $P(\mathcal{S})$ is e.g. a semi-Markov prior of the form (13).

Noninformative: A related approach sets

$$P(\mathcal{S}, \mathcal{I}) \propto P(\mathcal{S})$$

Note that this is equivalent to multiplying (29) by a factor $c(\mathcal{S})$, and strongly favors segment types which interact, as discussed in Section 9.2.

Conditionally informative: Alternatively, a non-uniform conditional prior on interactions may also be used in combination with segmentation priors:

$$P(\mathcal{S}, \mathcal{I}) \propto P(\mathcal{S})P(\mathcal{I} \mid \mathcal{S})$$

Some care is required in specifying $P(\mathcal{I} \mid \mathcal{S})$; examples of this approach are given in Section 9.2.

General informative Alternatively, general priors on interacting segmentations may be specified. By analogy to (27) for example, we might write

$$P(\mathcal{S}, \mathcal{I}) = P(\mathcal{S}) = P(\%H, \%E, \%L \in T, \#I \in \mathcal{I} \mid m)P(m) \tag{30}$$

in order to model the observed frequency of occurrence of various numbers of various types of β-sheets.

Section 9.2 and (Schmidler et al., 2004) provide concrete examples of some of these forms, including effects on predictive accuracy for applications in protein sequence analysis.

6 Inference and prediction in SSIMs

In this section we discuss inference and prediction with SSMs and SSIMs. Sections 7 and 8 discuss algorithmic issues associated with these tasks. Where appropriate, we use as an example the problem of protein secondary structure prediction.

6.1 Segmentation with interactions

We refer to the task of recovering the unobserved $(\mathcal{S}, \mathcal{I})$ given an observed sequence X as *segmentation* of X. Many applications involve segmentation problems, including parsing of human speech, identification of gene structure in DNA, prediction of protein secondary structure, and identification of change points in time series data.

We first consider SSMs without interactions. A commonly used estimate of $\mathcal{S}$ is the *maximum a posteriori* (MAP) value of $\mathcal{S}$, or *MAP segmentation*:

$$\begin{aligned} \mathcal{S}_{MAP} &= \arg\max_{\mathcal{S}} P(\mathcal{S} \mid X, \theta, \gamma) \\ &= \arg \max_{(m, \ell_1, \ldots, \ell_m, T_1, \ldots, T_m)} P(m, (\ell_i, T_i)_{i=1}^m \mid X, \theta, \gamma) \end{aligned} \tag{31}$$

The estimator $\mathcal{S}_{MAP}$ minimizes the *Bayesian expected loss* (Berger, 1985):

$$E_{\mathcal{S}|X}(L, \hat{\mathcal{S}}) = \sum_{\mathcal{S}} P(\mathcal{S} \mid X, \theta, \gamma) L(\mathcal{S}, \hat{\mathcal{S}})$$

for *0-1 loss* ($L(\mathcal{S}, \hat{\mathcal{S}}) = 0$ for $\mathcal{S} = \hat{\mathcal{S}}$, and 1 otherwise), making it the optimal estimator or *Bayes rule* for $\mathcal{S}$ in

the sense of minimizing the Bayes risk (Berger, 1985). An alternative estimator for $\mathcal{S}$ is the *marginal mode* predictor, known as "smoothing" in HMMs:

$$\mathcal{S}_{MM} = \{\arg\max_{T} P(T_{X_{[i]}} \mid X, \theta)\}_{i=1}^{n} \tag{32}$$

where $P(T_{X_{[i]}} \mid X, \theta)$ denotes the marginal posterior distribution over segment types at position i in the sequence, marginalized over all possible segmentations:

$$P(T_{X_{[i]}} \mid X, \theta) = \sum_{\mathcal{S}} P(\mathcal{S} \mid X, \theta) \mathbf{1}_{\{T_{X_i} = t\}} \tag{33}$$

and $\mathcal{S}_{MM}$ is the sequence of marginal modes at each position. $\mathcal{S}_{MM}$ is optimal under a *Hamming distance* loss $L(\mathcal{S}, \mathcal{S}^*) = \sum_{i=1}^{n} \mathbf{1}_{\{T_{X_i} = T^*_{X_i}\}}$. Thus if we wish to maximize the number of correctly classified positions then $\mathcal{S}_{MM}$ is preferred. A disadvantage of $\mathcal{S}_{MM}$ is that the resulting segmentation may have posterior mass zero. Section 9 demonstrates a significant improvement of $\mathcal{S}_{MM}$ over $\mathcal{S}_{MAP}$ in predicting protein secondary structure.

Segmentation under the framework of SSIMs introduced in Section 3.3 proceeds similarly. In a direct parallel to (31,32), we define

$$(\mathcal{S}, \mathcal{I})_{MAP} = \arg\max_{(\mathcal{S}, \mathcal{I})} P(\mathcal{S}, \mathcal{I} \mid X) \tag{34}$$

$$\mathcal{S}^{\mathcal{I}}_{MM} = \{\arg\max_{T} P(T_{X_{[i]}} \mid X)\}_{i=1}^{n} \tag{35}$$

where $P(T_{X_{[i]}} \mid X)$ is now marginalized over interacting segmentations:

$$P(T_{X_{[i]}} \mid X) = \sum_{(\mathcal{S}, \mathcal{I})} P(\mathcal{S}, \mathcal{I} \mid X) \mathbf{1}_{\{T_{X_i} = t\}} \tag{36}$$

Again, these predictors are Bayes optimal under *0-1* and Hamming distance loss, respectively. However use of uniform priors (28) yields a marginal prior $P(\mathcal{S}^{\mathcal{I}})$ which is highly biased towards interacting segments, and thus is expected to perform poorly in prediction of marginal quantities via (35). Conversely, conditional priors (29) significantly downweight the contribution of any particular interaction and are unlikely to improve marginal predictions through incorporation of joint-segment information. Instead, we focus on prediction of the interactions themselves.

6.2 Contact map prediction

In addition to providing segment locations and types, the MAP interacting segmentation $(\mathcal{S},\mathcal{I})_{MAP}$ defined by (34) also includes the MAP set of segment interactions and associated parameters. For the β-sheet models described in Section 9.2 this provides MAP β-sheet topologies.

We are often interested in predicting interactions between specific sequence positions, which we summarize by a *contact map* matrix $C_{n\times n}$ where

$$C_{ij} = \begin{cases} 1 & \text{If } X_i \text{ and } X_j \text{ are paired in a segment interaction} \\ 0 & \text{otherwise} \end{cases}$$

We denote by C^{MAP} the contact map estimator derived from $(\mathcal{S},\mathcal{I})_{MAP}$, so $C^{MAP}_{ij} = 1$ if X_i and X_j are paired in $(\mathcal{S},\mathcal{I})_{MAP}$. We define the *marginal predicted contact map* to be the matrix C^{MM} where

$$C^{MM}_{ij} = P(i \leftrightarrow j) = \sum_{\mathcal{S},\mathcal{I}} P(\mathcal{S},\mathcal{I} \mid X)\mathbf{1}_{\{i\leftrightarrow j\}} \tag{37}$$

are the marginal probabilities of each potential contact marginalized over all possible segmentations *and* segment interactions. C^{MM} therefore yields a Bayes estimator for C under Hamming loss.

Calculation of the quantities $(\mathcal{S},\mathcal{I})_{MAP}$, $\mathcal{S}^{\mathcal{I}}_{MM}$, C^{MAP}, and C^{MM} defined here for SSIMs is significantly more difficult than the computations in SSMs, and will be discussed next. Experiments comparing C^{MAP} and C^{MM} as predictors of true contacts in protein sequences are described in Section 9.

7 Dynamic Programming Algorithms for SSMs

We now turn to issues of computation with SSMs. We give efficient algorithms for several types of SSMs using dynamic programming. Section 8 discusses computation in SSIMs.

Prediction and inference with SSMs requires calculation of posterior quantities such as (31) and (32) given an observed sequence X. The algorithms of choice depend on the form of the joint distribution (2), including the prior. We consider three cases defined in Section 5.1: segment-decomposable priors; conditionally segment-decomposable priors; and non-decomposable priors.

7.1 Segment-decomposable priors

Calculations for SSMs are most efficient (barring special cases such as HMMs) for segment-decomposable priors, such as uniform priors (22,26), or semi-Markov process priors with $P(m) \propto 1$ improper (24,13), which lead to a joint distribution which is also segment-decomposable. Exact calculations may be done using standard HMM-type forward-backward and Viterbi algorithms generalized to HSMMs (Rabiner, 1989; Stormo and Haussler, 1994; Schmidler et al., 2000). In particular, the $\mathcal{S}_{MAP}$ may be calculated by dynamic programming using forward variables:

$$\delta(j,t) = \max_{\substack{v<j \\ l \in \mathcal{A}_T}} [\delta(v,l) f(e_- = v, e = j, T_- = l, T = t)] \tag{38}$$

in a procedure analogous to the Viterbi algorithm for HMMs (Rabiner, 1989). Here e_- represents the endpoint of the previous segment. For the SSM of (11) and Section 9.1 we have

$$f(v,j,l,t) = P(X_{[v+1:j]} \mid T = t) P(\ell = j - v \mid T = t) P(T = t \mid T_- = l)$$

The algorithm recursively calculates $\delta(j,t)$ for $j = 1, \ldots, n$ and $t \in \mathcal{A}_T$, then reconstructs the MAP segmentation by setting $e_m^* = n$, $T_m^* = \arg\max_{l \in \mathcal{A}_T} \delta(n,l)$, and tracing backwards. This calculation requires $O(n^3)$ steps. Often a maximum segment length D may be imposed, so the maximization (38) begins at $v = j - D$ and the algorithm becomes $O(nD^2)$, linear in n. Certain segment models permit further reduction to $O(nD)$, but this does not hold in general. Figure 4a shows that $D = 30$ is sufficient to account for nearly all protein structural segments; experiments in Section 9 use this value.

The marginal posterior distributions (33) may also be calculated efficiently, using forward/backward variables α/β:

$$\alpha(j,t) = \sum_{v<j} \sum_{l \in \mathcal{A}_T} \alpha(v,l) f(v,j,l,t) \tag{39}$$

$$\beta(j,t) = \sum_{v>j} \sum_{l \in \mathcal{A}_T} \beta(v,l) f(j,v,t,l) \tag{40}$$

$$P(T_{X_i} = t \mid X, \theta) = \frac{1}{Z} \sum_{j<i} \sum_{k \geq i} \sum_{l \in \mathcal{A}_T} \alpha(j,l) \beta(k,t) f(j,k,l,t) \tag{41}$$

where Z is the marginal likelihood $P(X \mid \theta)$, available directly from the forward pass (39). Calculation of (39) and (40) requires $O(n^3)$ (or $O(nD^2)$) steps, while (41) yields the marginal posterior distribution at each

position in $O(n^2)$ (or $O(nD)$) using:

$$P(T_{X_{[i+1]}} = t \mid X, \theta) = P(T_{X_{[i]}} = t \mid X, \theta) + \frac{1}{Z}\left[\sum_{k>i}\sum_{l\in\mathcal{A}_T}[\alpha(i,l)\beta(i+1,k)f(i,k,l,t)] - \sum_{j<i}\sum_{l\in\mathcal{A}_T}[\alpha(j,l)\beta(i,t)f(j,i,l,t)]\right] \quad (42)$$

Thus using the algorithms given in this section, both predictors $\mathcal{S}_{MAP}$ and $\mathcal{S}_{MM}$ may be calculated efficiently under segment-decomposable priors.

7.2 Conditionally segment-decomposable priors

Another important case occurs when $P(\mathcal{S})$ (and thus $P(\mathcal{S} \mid X)$) is conditionally segment-decomposable given m the number of segments but $g(m)$ itself is not segment-decomposable. Examples include:

(i) Semi-Markov process priors (24) with non-uniform marginal prior $P(m)$, where $g(m \mid \theta) = P(m)$ and $f()$ as above.

(ii) Sequence dependent priors with $P(\mathcal{S} \mid n) = h(m,n)$ such as (25), where $g(m \mid n, \theta) = h(m,n)$ and $f(v,j,t,l \mid m,n) = P(X_{[v+1:j]} \mid T = t)$.

Here the algorithms of the previous section do not apply, and must be adapted. Algorithms for this case may be adapted from (Auger and Lawrence, 1989) using the following forward and backward variables:

$$\delta(j,t,k) = \max_{\substack{v<j \\ l\in\mathcal{A}_T}}[\delta(v,l,k-1)f(v,j,l,t)]$$

$$\alpha(j,t,k) = \sum_{v<j}\sum_{l\in\mathcal{A}_T}\alpha(v,l,k-1)f(v,j,l,t)$$

$$\beta(j,t,k) = \sum_{v>j}\sum_{l\in\mathcal{A}_T}\beta(v,l,k-1)f(j,v,t,l)$$

defined for $k = 1,\ldots,n$. Now $\mathcal{S}_{MAP}$ is reconstructed by setting

$$(m*, T^*_{m^*}) = \arg\max_{\substack{m\in\{1,\ldots,n\} \\ l\in\mathcal{A}_T}}\delta(n,l,m)g(m)$$

and $e^*_{m^*} = n$ and tracing back recursively. These computations require $O(n^4)$ (or $O(n^2D^2)$) steps, a factor of n slower than in the previous section. Computation of the marginal posterior distributions $P(T_{X_{[i]}} \mid X)$

requires marginalization over m:

$$P(T_{X_i} = t \mid X, \theta) = \frac{1}{Z} \sum_{m=1}^{n} g(m) \sum_{q=1}^{m} \sum_{j<i} \sum_{k \geq i} \sum_{l \in \mathcal{A}_T} \alpha(j, l, q-1) \beta(k, t, m-q+1) f(j, k, l, t) \quad (43)$$

Thus we have efficient algorithms for calculation of posterior quantities such as $\mathcal{S}_{MAP}$ and $\mathcal{S}_{MM}$ in SSMs with conditionally segment-decomposable priors.

7.3 Non-decomposable priors

In the case of general non-decomposable priors of the form (27), efficient algorithms do not exist. However we may obtain approximate inference using MCMC techniques similar to those provided in Section 8 for SSIMs.

8 MCMC Algorithms for SSIMs

We now turn to posterior calculations with SSIMs. Unlike SSMs, efficient algorithms do not exist for SSIMs in the general case, where calculation of posterior quantities such as $(\mathcal{S}, \mathcal{I})_{MAP}$, $\mathcal{S}^{\mathcal{I}}_{MM}$, C^{MAP}, C^{MM}, or marginal likelihood $P(X)$ for SSIMs is a hard computational problem. SSIMs violate precisely the conditional independence structure of SSMs (2) that is critical for recursive decomposition of the joint distribution $P(X, \mathcal{S})$ to enable dynamic programming solutions. In general, introduction of joint-segment models into SSMs makes exact calculation of posterior probabilities intractable. Instead we describe Markov chain Monte Carlo algorithms for inference and prediction with general SSIMs.

8.1 Markov chain Monte Carlo segmentation

Markov chain Monte Carlo (MCMC) is now a standard tool for Bayesian inference with complex models, see e.g. (Gilks et al., 1996). We briefly describe a reversible-jump MCMC algorithm (Green, 1995) for inference with SSIMs.

MCMC for SSMs We begin with an MCMC algorithm for sampling from SSM joint distributions (2); we then extend this to SSIMs (4). To construct a Markov chain on the space of segmentations, we combine

Gibbs sampling steps:

$$T_k^{(i+1)} \sim P(T_k \mid \ell_k^{(i)}, S_{[-k]}^{(i)}) \quad \text{and} \quad \ell_k^{(i+1)} \sim P(\ell_k \mid T_k^{(i)}, S_{[-k]}^{(i)})$$

with Metropolis proposals from $\mathcal{S}$ to $\mathcal{S}^*$:

- *Segment split*: Split S_k into two segments (S_{k^*}, S_{k^*+1}) with $m^* = m+1$, $e_{k^*+1} = e_k$, $e_{k^*} \sim U[s_k, e_k - 1]$, and with probability $\frac{1}{2}$ set $T_{k^*} = T_k$ and $T_{k^*+1} = T_{new} \sim U[\mathcal{A}_T]$; 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 $m^* = m - 1$.

The factorization of (2) makes calculation of exact conditionals efficient, involving only terms local to the affected segment:

$$P(T_k = t \mid \ell_k, S_{[-k]}) \propto$$
$$P(T_k = t \mid T_{k-1}) P(\ell_k \mid T_k = t) P(X_{[s_k:e_k]} \mid T_k = t) P(T_{k+1} \mid T_k = t)$$

Split and *merge* moves change model dimension, and are accepted according to a reversible-jump Metropolis criteria, again involving only local terms:

$$\rho(\mathcal{S}, \mathcal{S}^*) = \left[\frac{\prod_{j=k}^{k+1} P(X_{[s_j^*:e_j^*]} \mid T_j^*) P(\ell_j^* \mid T_j^*) P(T_{j+1}^* \mid T_j^*)}{P(X_{[s_k:e_k]} \mid T_k) P(\ell_k \mid T_k) P(T_{k+1} \mid T_k)} \right] \left[\frac{m(\ell_k - 1)|\mathcal{A}_T|}{(m+1)} \right]$$

Together these 4 steps are sufficient to yield an ergodic Markov chain, and hence an algorithm for MCMC segmentation under SSM models. However two additional moves facilitate rapid mixing of the Markov chain:

- *Segment create*: Given $\mathcal{S}$, propose $\mathcal{S}^*$ with $m^* = m + 2$ segments by splitting segment k into three segments $(k, k+1, k+2)$, as follows: Set $e_{k+2}^* = e_k$, draw $l_1 \neq l_2 \sim U[s_k, e_k - 1]$, and set $e_k^* = \min(l_1, l_2)$, $e_{k+1}^* = \max(l_1, l_2)$, $T_k^* = T_{k+2}^* = T_k$, and draw $T_{k+1}^* \sim U[\mathcal{A}_T \backslash \{T_k\}]$.
- *Segment remove*: Similar to *segment create*, but segment k with $T_{k+1} = T_{k-1}$ is removed and its immediate neighbors merged, yielding $m^* = m - 2$.

Acceptance ratios for these dimension-altering Metropolis moves are similar to those above. The result is a faster mixing Markov chain (see Figure 4b).

MCMC for SSIMs For sampling from SSIM models (4), the above algorithm must be supplemented by additional moves involving segment interactions:

- *Segment join*: Given $(\mathcal{S}, \mathcal{I})$, propose an interaction I^* between two segments $S_j, S_k \notin \mathcal{I}$ and set $\mathcal{I}^* = \mathcal{I} \cup \{I^*\}$.
- *Segment split*: Reverse of *segment join*; splits a 2-segment interaction into two independent segments.
- *Segment insert* and *Segment delete* Insert a segment $S_i \notin \mathcal{I}$ into an existing interaction $I \in \mathcal{I}$, or remove a segment from an interaction.
- *Segment align*: Given $I_j \in \mathcal{I}$, sample interaction parameters H_j.

These moves follow the general format of those described previously, except for *align* which is application-specific. Section 9.2 details these moves for an application to protein β-sheet prediction.

This combined set of moves provides an MCMC scheme for inference with SSIMs, enabling approximate calculation of posterior quantities $P(\mathcal{S}, \mathcal{I} \mid X)$ such as $Struct_{MAP}$ and $Struct_{MM}$.

9 SSIM models for protein sequence analysis

Protein structure prediction was described briefly as a motivating problem for SSIMs in Section 2.

9.1 SSMs models for secondary structure prediction

Protein secondary structure prediction is the problem of assigning each sequence position to one of the classes α-helix, β-strand, or loop/coil. SSMs have been applied to this problem by developing segment models for each class, yielding results comparable to the best published (Schmidler et al., 2000). Segment models take the form:

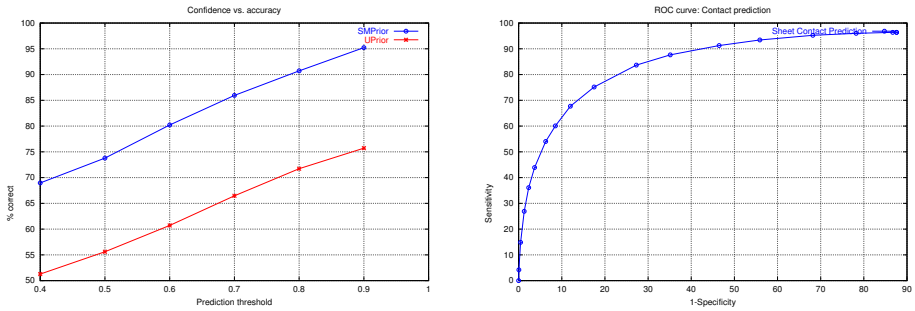
$$P(X_{[s_j:e_j]} \mid T_j = H) = \prod_{i=s_j}^{e_j} P_{M(i)}\left(X_i \mid X_{[s_j:i-1]} \mid T_j = H\right)$$

where

$$M(i) = \begin{cases} N_{i-s_j} & i - s_j \le \ell_N^H \\ C_{e_j - i} & e_j - i \le \ell_C^H \\ I & \text{otherwise} \end{cases}$$

	Total Q_3	Helix Q_α^{obs} (Q_α^{pred})	Strand Q_β^{obs} (Q_β^{pred})	Loop Q_L^{obs} (Q_L^{pred})	Helix C_α	Strand C_β	Loop C_L
MM_{SM}	68.0	65.1 (68.1)	44.9 (58.7)	79.2 (70.6)	.53	.41	.46
MAP_{SM}	64.0	69.7 (59.9)	24.4 (62.2)	76.7 (66.6)	.48	.31	.38
MM_U	50.3	58.7 (53.3)	71.2 (33.4)	37.0 (78.5)	.36	.28	.30
MAP_U	50.2	63.2 (48.1)	58.5 (35.0)	39.2 (72.1)	.33	.26	.27
SSIM	65.1	48.5 (76.3)	59.6 (46.7)	77.0 (69.9)	.49	.39	.44
SSM_{MC}	67.9	64.7 (68.1)	44.7 (58.4)	79.1 (70.4)	.53	.41	.46

Table 1: Cross-validation results for protein secondary structure prediction using SSMs, comparing $\mathcal{S}_{MAP}$ and $\mathcal{S}_{MM}$ predictors under semi-Markov (SM) and uniform (U) priors. Q_3 denotes % correct over all 3 states; $Q_\alpha^{obs/pred}$ denotes sensitivity/positive predictive value and C_α the Matthews correlation coefficient (Matthews, 1975) for class α. Also shown: SSIM model, and SSM model using MCMC inference (SSM_{MC}) in place of exact algorithms.



(a) (b)

Figure 5: (a) Strong correlation between predictive accuracy and predicted probability, shown for both semi-Markov and uniform priors. (b) ROC curve for β-sheet contact predictions as a function of (log) probability of predicted contact.

indexes a collection of position-specific conditional distributions P_{N_i/C_i} for the first $\ell_{N/C}$ N-/C-terminal amino acids (N_i denotes the i^{th} position from the N-terminus), and an identically distributed but dependent distribution I for internal positions.

Table 1 and Figure 5a show the results of applying this SSM model to secondary structure prediction on a large non-redundant database of experimentally determined protein structures via cross-validation. The S_{MM} predictor consistently out performs S_{Map} in correctly predicted positions. We note that (41) provides the *exact* marginal posterior distribution over segment types at each position, averaging over *all possible* segmentations, and hence provides a clear indication of prediction uncertainty at each position. Figure 5a shows that this measure correlates strongly with prediction accuracy. Results are shown using both semi-Markov (13) and uniform (22) segmentation priors, demonstrating that maximum likelihood segmentation is a poor choice in this setting.

9.2 SSIM models for β-sheet contact maps

We now consider SSIMs for protein β-sheet prediction. As described in Section 2, β-sheets are an example of non-local interactions in protein sequences which give rise to long-range sequence correlations, violating the conditional independence assumptions of SSMs. To model these non-local dependencies, we have developed SSIMs involving joint-segment models neighboring β-strands within β-sheets, building on previous work using empirical pair potentials (Hubbard, 1994). Analysis of inter-strand side chain dependencies and more detailed model development for this application will be reported elsewhere (Schmidler et al., 2004).

Simple joint-segment models for β-sheets are given by

$$P(\{R_{[s_{H_j}:e_{H_j}]}\}_{j=1}^{k} \mid \mathcal{S}, \mathcal{I}) = \left[\prod_{j=1}^{k} \prod_{i=s_{H_j}}^{e_{H_j}} P(R_{[i]})\right] \prod_{j=1}^{k-1} \prod_{i=0}^{l_{jj+1}-1} \frac{P(R_{[n_{jr}+i]}, R_{[p(i,A)]})}{P(R_{[n_{jr}+i]})P(R_{[p(i,A)]})} \quad (44)$$

where conditioning on $(\mathcal{S}, \mathcal{I})$ has been suppressed for clarity, and H_j is the segment index for the j^{th} β-strand, n_{jr} denotes the first N-terminal position of S_{H_j} interacting with the right neighbor, l_{jj+1} the number of positions interacting between β-strand S_{H_j} and neighboring strand $S_{H_{j+1}}$, and A denotes the orientation of pairing (parallel or anti-parallel), with

$$p(i, A) = \begin{cases} c_{j+1l} - i & \text{if} \quad A = -1 \quad \text{(pairing is anti-parallel)} \\ n_{j+1l} + i & \text{if} \quad A = 1 \quad \text{(pairing is parallel)} \end{cases}$$

giving the right neighbor residue in the i^{th} pair. This model incorporates a simple dependency between amino acids within a β-sheet: each cross-strand pair is *iid*, with no intra-segment dependency.

The prior distribution on β-sheet interactions is given by:

$$P(\mathcal{I} \mid \mathcal{S}) = P(p \mid \mathcal{S}) \prod_{j=1}^{p} P(I_j \mid \mathcal{S})$$

$$P(I_j \mid \mathcal{S}) = \frac{1}{k_j!} P(A) \prod_{i=1}^{k_j - 1} P(n_{ir}) P(c_{ir}) P(n_{i+1l}) P(c_{i+1l}) \quad (45)$$

where $P(A)$ is the prior probability of parallel vs. anti-parallel β-sheets estimated from database frequencies, and we currently take $P(p \mid \mathcal{S}) \propto 1$. Under (45) all strand topologies are equally likely, and registration

parameters $(n_{il/r}, c_{il/r})$ are given priors $P(n_{il}) = \frac{1}{3}$ for $(n_{il} - s_i) \leq 2$, 0 otherwise, making the the first (last) interaction or hydrogen bond between two paired strands uniform over the first (last) 3 positions.

In this context, the MCMC moves from Section 8 propose β-sheet formation by joining unpaired β-strands, destruction by separating paired strands, and insertion/deletion of β-strands into existing β-sheets during sampling. *Segment align* samples the registration of a β-strand with its neighbor. $(\mathcal{S}, \mathcal{I})_{MAP}$ provides a MAP set of β-sheets including number, topology, orientation, and registration parameters. $\mathcal{S}^{\mathcal{I}}_{MM}$ provides predictions of secondary structure which marginalize over possible β-sheet formation, allowing prediction of secondary structure and β-sheet topology in within a common statistical framework.

Experimental results: Figure 6 shows predictions for two small proteins, BPTI and (5pti) and ribonuclease A. While the native contacts in these small proteins are typically among the high probability predicted contact regions, at least one alternative pairing for each strand commonly exists.

Figure 5 shows an ROC curve obtained by applying this model to a random sample of 100 proteins from the non-redundant set of Section 9.1. Further development of the β-sheet models to improve predictive accuracy is ongoing (Schmidler et al., 2004).

Table 1 shows the results of using the marginals of the SSIM model for structure prediction. It can be seen that this results in a small decrease in accuracy, resulting from a higher sensitivity (but lower specificity) in predicting β-strand positions. This is attributable to the asymmetric treatment of segment types in the current SSIM model, since only β-strands are permitted to participate in interactions. A small amount of error is also introduced by using the Monte Carlo approximation, as demonstrated by the results shown for prediction under the SSM model via the MCMC algorithm.

10 Discussion

We have discussed stochastic segment models (SSMs) and algorithms for analysis of sequence data, and introduced a new class of models for long-range dependence called stochastic segment interaction models (SSIMs). We describe priors for both classes of models. Examples were given for the use of SSMs and SSIMs for prediction of protein secondary structure and protein β-sheet contact maps from sequence, respectively.

The SSIM framework developed here is quite general, and applicable to any sequential data involving long-range dependency between sequentially separate blocks. Other applications which exhibit such repeated patterns may included econometrics and finance, computer intrusion detection, or other physical processes.

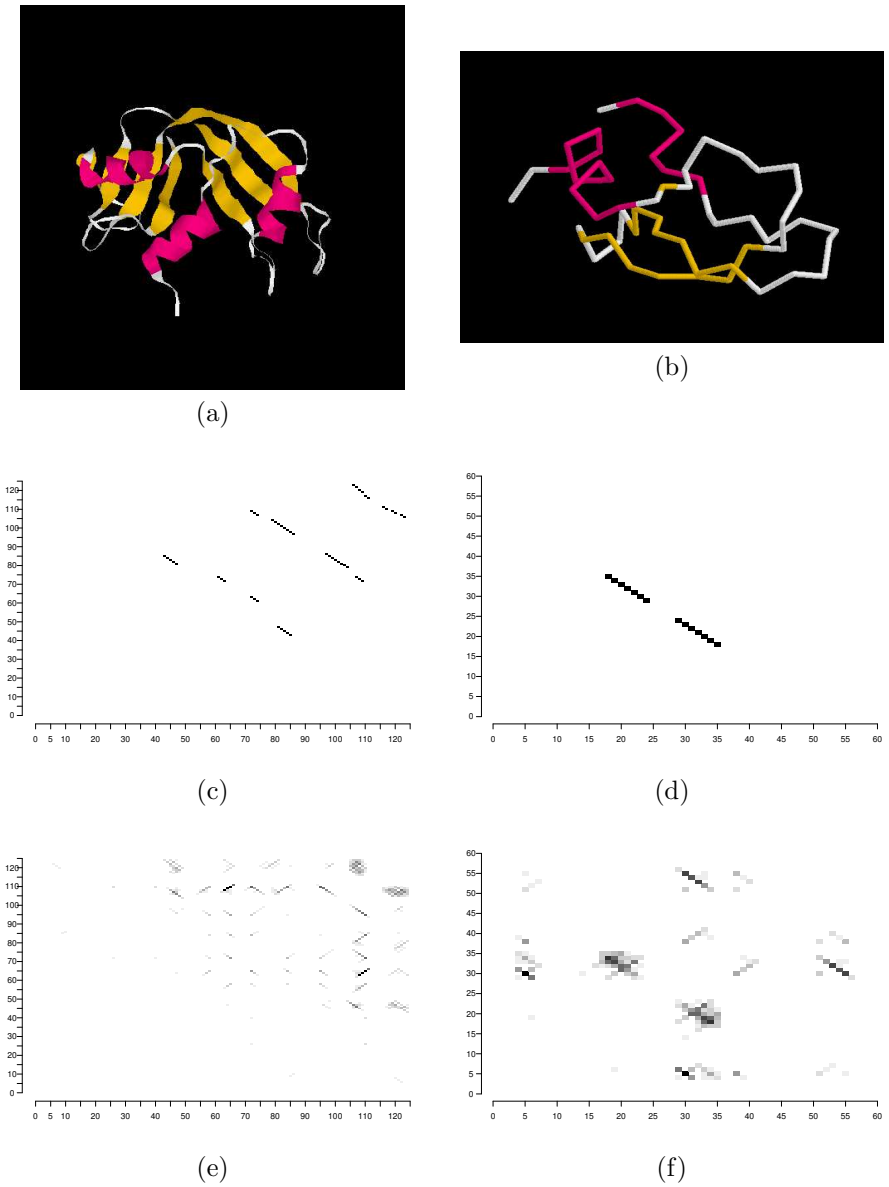


Figure 6: β-sheet contact map prediction for bovine pancreatic trypsin inhibitor (5pti) and ribonuclease A (1rbx) proteins. Shown are (a,b) X-ray crystallographic structure obtained from Protein Data Bank, (c,d) true contact map derived from crystal structure, and (e,f) probability of contacts predicted from sequence. Contact map axes represent position in sequence. Shading of pixels (x,y) indicates predicted probability of residues x,y forming contacts within a β-sheet.

References

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.

Auger, I. E. and Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.*, 51(1):39–54.

Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. A. (1994). Hidden Markov models of biological primary sequence information. *PNAS*, 91:1059–1063.

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *J. Amer. Statist. Assoc.*, 88:309–319.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, 2^{nd} edition.

Braun, J. V. and Muller, H.-G. (1998). Statistical methods for DNA sequence segmentation. *Stat. Sci.*, 13:142–162.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol. Biol.*, 268:78–94.

Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.*, 51(1):79–94.

Depmster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39:1–38.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Eddy, S. and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucl. Acids Res.*, 22(11):2079–2088.

Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.*, 6:361–365.

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.

Hopcroft, J. E. and Ullman, J. D. (1979). *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley.

Hubbard, T. J. P. (1994). Use of β-strand interaction pseudo-potentials in protein structure prediction and modeling. In Lathrop, R. H., editor, *Biotechnology Computing Track, 27th HICSS*, pages 336–354. IEEE Computer Society Press.

Kindermann, R. and Snell, J. L. (1980). *Markov Random Fields and Their Applications*. American Mathematical Society.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *J Mol. Biol.*, 235:1501–1531.

Kulp, D., Haussler, D., and Reese, M. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. In States, D. J., Agarwal, P., Gaasterland, T., Hunter, L., and Smith, R., editors, *Proceedings, Fourth International Conference on Intelligent Systems in Molecular Biology*, pages 134–142. AAAI Press.

Lari, K. and Young, S. J. (1990). The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56.

Lari, K. and Young, S. J. (1991). Applications of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 5:237–57.

Levinson, S. E. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Comp. Speech Lang.*, 1:29–45.

Liu, J. S. and Lawrence, C. E. (1996). Unified Gibbs method for biological sequence analysis. In *Amer. Statist. Assoc., Statist. Comp. Section*.

Liu, J. S. and Lawrence, C. E. (1999). Bayesian inference on biopolymer models. *Bioinformatics*, 15(1):38–52.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman & Hall.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, 405:442–451.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:11–05–1110.

Ostendorf, M., Digalakis, V., and Kimball, O. A. (1996). From HMMs to segment models: A unified view of stochastic modeling for speech recognition. *IEEESAP*, 4:360–378.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286.

Russell, M. J. and Moore, R. K. (1985). Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition. In *ICASSP*, pages 5–8, Tampa, FL.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjolander, K., Underwood, R. C., and Haussler, D. (1994). Stochastic context-free grammars for tRNA modeling. *Nucl. Acids Res.*, 22(23):5112–5120.

Schmidler, S. C. (2002). *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.

Schmidler, S. C., Liu, J. S., and Brutlag, D. L. (2001). Bayesian protein structure prediction. In *Case Studies in Bayesian Statistics*, volume 5, pages 363–378.

Schmidler, S. C., Liu, J. S., and Brutlag, D. L. (2004). Bayesian modeling of non-local interactions in protein sequences. (*submitted*).

Searls, D. B. (1993). The computational linguistics of biological sequences. In Hunter, L., editor, *Artificial Intelligence and Molecular Biology*, pages 47–120. MIT Press.

Snyder, E. E. and Stormo, G. D. (1993). Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks. *Nucl. Acids Res.*, 21(3):607–613.

Stephens, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *Appl. Statist.*, 43(1):159–178.

Stormo, G. D. and Haussler, D. (1994). Optimally parsing a sequence into different classes based on multiple types of evidence. In Altman, R., Brutlag, D., Karp, P., Lathrop, R., and Searls, D., editors, *Proceedings,*

Second International Conference on Intelligent Systems in Molecular Biology, pages 369–375. AAAI Press.

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

Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52.

Zuker, M. and Sankoff, D. (1984). RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46(4):591–621.

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of larger RNA sequences using thermodyamics and auxiliary information. *Nucl. Acids Res.*, 9:133–148.