STATISTICAL MODELS AND MONTE CARLO METHODS FOR PROTEIN STRUCTURE PREDICTION

A DISSERTATION
SUBMITTED TO THE PROGRAM IN BIOMEDICAL INFORMATICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Scott C. Schmidler
May 2002

© Copyright by Scott C. Schmidler 2002
All Rights Reserved

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Douglas L. Brutlag
(Biochemistry)
(Principal Advisor)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jun S. Liu
(Statistics, Harvard University)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Russ B. Altman
(Medicine)

Approved for the University Committee on Graduate Studies:

Abstract

As we enter the post-genome era, widespread availability of genomic data promises to revolutionize biomedicine, providing fundamental insights into the molecular mechanisms of disease and pointing the way to developing novel therapies. However important hurdles remain, including understanding the function and mechanism for the proteins encoded by genomic sequences. While function and mechanism are dictated by a protein's native structure, prediction of protein structure from sequence remains a difficult unsolved problem.

In this dissertation, I develop a novel framework for protein structure prediction from amino acid sequence, based on a new class of generalized stochastic models for sequence/ structure relationships. I introduce a formal Bayesian framework for synthesizing the varied sources of sequence information in structure prediction using joint sequence-structure probability models based on structural segments. I describe a set of probabilistic models for structural segments characterized by conditional independence of inter-segment positions, develop efficient algorithms for prediction in this class of models, and evaluate this approach via cross-validation experiments on experimental structures. This approach yields secondary structure prediction accuracies comparable to the best published methods, and provides accurate estimates of prediction uncertainty, allowing identification of regions of a protein predicted at even higher accuracies.

I then generalize this Bayesian framework to models of the non-local interactions in protein sequences involved in tertiary folding. I develop Monte Carlo algorithms for inference in this class of models, and demonstrate this approach with models for correlated mutations in β-sheets. Case studies and cross-validation experiments

demonstrate this approach for predicting β-strand contact maps, providing important information about protein tertiary structure from sequence alone.

This dissertation provides a suite of statistical models and computational tools for protein structure prediction. In addition, the models developed here generalize existing stochastic models in important ways. I relate these new models to existing generalized hidden Markov and stochastic segment models, showing the latter to be special cases of the former. Further, the interaction models developed here represent a novel class of stochastic models for sequences of random variables with complex long-range dependency structure. These new models, and the associated algorithms, are likely to be of broader statistical interest.

Acknowledgments

Anyone who has written a Ph.D. dissertation knows that it is not done alone. This document would not exist without the contributions of many people, beginning long before my arrival at Stanford and continuing after my departure.

During my time at Stanford, I have been extraordinarily fortunate to have the guidance and friendship of not one but two outstanding thesis advisors, Doug Brutlag and Jun Liu. Their excitement and enthusiasm about science, along with their patience and direction, has made the trip an enjoyable and rewarding one.

Russ Altman brought me to Stanford in the first place, and I have benefited from his tremendous energy, frank advice, and thankless support throughout my graduate career. Trevor Hastie provided insight and expertise on numerous ideas and occasions. Stuart Russell introduced me to research as a Berkeley undergrad, and opened my mind to the possibility of a career in science; without Stuart I would never have begun.

My research has benefited enormously from the diverse environment in which it took place. Faculty and students of Biochemistry, Statistics, and especially (Bio) Medical Informatics have provided many stimulating discussions and valuable advice. Particular thanks goes to my contemporaries in the Brutlag Bioinformatics group, including Amit Singh, Tom Wu, Craig Nevill-Manning, Steve Bennett, and Jimmy Huang, who made the lab an enjoyable and educational place to spend my years. My research was supported by the National Library of Medicine and the National Center for Human Genome Research.

Finally, my heartfelt thanks to my parents and family, who have supported and encouraged me in everything I have done, and who provided the foundation and

opportunities that brought me here; to friends and fellow travelers Scott Pegg and Ramon Felciano, who reminded me that a passion for science does not preclude a passion for life; and to Gillian Sanders, my best friend and lover, whose unwavering confidence in me was perhaps my greatest strength. I thank her for her love and support, for her patience and resilience, for her late-night secretarial skills, and for introducing me to climbing, while having the good sense to wait until I was (nearly) finished. I could not have done this without her.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Structural bioinformatics in the post-genome era

The Human Genome Project is nearing completion, with rough-drafts currently in the public and private domain (Consortium, 2001; Venter et al., 2001). Many complete genomes for pathogenic organisms have been available for some time (Fleischmann et al., 1995; Fraser et al., 1995; Tomb et al., 1997), with many more under way. Widespread availability of this data promises to revolutionize 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. Unfortunately, while the sequence of the 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 currently estimated to

contain approximately 35,000 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 known as "structural bioinformatics", a sub-field of the emerging discipline of bioinformatics. It is to this field that the majority of the developments in this dissertation belong.

Structural bioinformatics is not limited to protein structure prediction. Already looking beyond the completion of the Human Genome Project, efforts are emerging to develop and make public a large, structurally diverse set of high-resolution experimentally determined protein structures mapping the universe of naturally occurring folds (Burley et al., 1999; Montelione and Anderson, 1999). As discussed below, such developments will continue to magnify the importance of the work presented here. Moreover, realization of such a goal will provide a multitude of new computational and statistical problems with the potential to make enormous contributions to molecular biology. The wide range of scientifically important and theoretically challenging problems emerging makes it an exciting time for Bioinformatics as a field.

The problems addressed in this dissertation lie in the realm of protein structure prediction. Interest in protein structure prediction ranges far beyond Bioinformaticians, as practical solutions to this problem have implications for nearly all of biology and molecular medicine. In particular, accurate and interpretable models for protein structure prediction have implications throughout the protein science aspects of molecular biology, which encompass the study of protein chemistry and mechanism; the kinetics and thermodynamics of protein folding, including protein engineering; protein-protein interactions and signaling; and protein-ligand interactions, including nucleotide-binding proteins and regulation of gene expression as well as small-molecule binding and pharmaceutical development. The fundamental principles of protein folding lie at the core of each of these problems. The enormous variety of roles played by proteins in living organisms, among them structural, enzymatic, regulatory, and signaling, makes a rigorous understanding of this process all the more critical. Progress in these areas is expected to yield a greater understanding of the biochemical processes of life and disease than ever before.

The study of protein structure, folding, and function has a long and distinguished

history in the 20^{th} century. Since Linus Pauling first predicted theoretically the existence of α-helices and β-sheets in 1951 (Pauling et al., 1951; Pauling and Corey, 1951), predictions which were later verified experimentally and which greatly influenced the discovery of the double-helical structure of DNA (Watson and Crick, 1953), protein folding and its role in determining function has been the subject of intensive research effort. Still, the long sought-after goal of theoretical models for protein folding which provide practical algorithms for prediction of protein tertiary structure from sequence remains unrealized.

In this dissertation, I provide a fresh look at the problem of protein structure prediction from the perspective of probability theory and statistics. I develop a formal framework for synthesizing the varied sources of information about a protein's structure under a framework of Bayesian inference. Within this framework I develop a class of probabilistic models of structural segments which can be fit to experimental data, and demonstrate the practicality of this approach by applying it to the prediction of protein secondary structure and β-sheet topology from amino acid sequence. The methods developed in this dissertation are evaluated via computational experiments involving blind prediction on a database of experimentally determined protein structures and shown to be both accurate and informative. The interpretability of these models, and links with statistical mechanical theories of folding, make them novel tools for studying protein folding on a computer. In particular, they enable calculation of the contribution of various physicochemical properties of protein sequences to the accuracy of prediction, and allowing the models developed here to be extended to incorporate future advances in the understanding of protein folding.

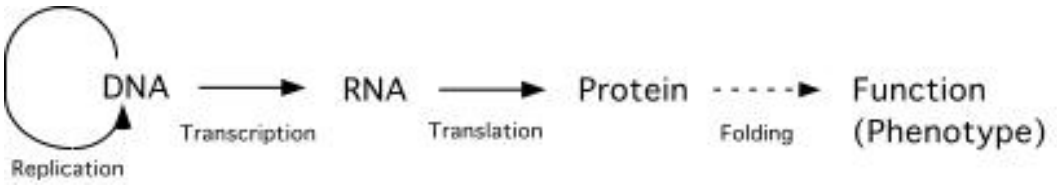


Figure 1.1: The *Central Dogma* of molecular biology. Genetic information is stored and replicated as DNA, transcribed to RNA, and translated to proteins. Proteins perform the vast majority of biochemistry required by living organisms.

1.2 Proteins and their structures

1.2.1 Role of protein structure in the flow of genetic information

To place the problem of protein structure prediction in context, I briefly review the "central dogma" of molecular biology, summarized in Figure 1.1. The genetic information of an individual organism is stored in its DNA, located in the nucleus of cells for eukaryotes such as humans. This DNA is transcribed into messenger RNA (mRNA) and transported out of the nucleus, where the (majority of) RNA is translated into protein sequence by ribosomes. Proteins are therefore the tangible product of expressing genes stored in the DNA. In actuality, this process of genetic expression is highly complex and carefully regulated, and the subject of numerous branches of scientific study (Stryer, 1995).

The primary concern of this dissertation is the protein sequence produced by this process of translation. The basic structure of a protein sequence is shown in Figure 2.1, and is discussed in more detail in Chapter 2. Proteins perform the vast majority of the biochemistry required by living organisms, playing various catalytic, structural, regulatory, and signaling roles required for cellular metabolism, development, differentiation, and replication. The key to the wide variety of functions exhibited by individual proteins is not the linear sequence as shown in Figure 2.1 however, but the three dimensional configuration adopted by this sequence in its native environment (shown in Figure 2.2). In order to understand protein function at the molecular level then, it is crucial to study the structure adopted by a particular

sequence. Unfortunately our understanding of the physical process by which a sequence achieves this structure, known as *protein folding*, remains inadequate despite decades of study. In particular, serious difficulties present themselves when one attempts to simulate this folding process computationally in order to learn the structure of a given protein sequence (Chapter 3).

1.2.2 Predicting structure from sequence

In the absence of practical methods for simulating protein folding, significant effort has shifted to the problem of protein structure *prediction*. In structure prediction, we abandon any attempt to simulate the actual physical process of folding, and instead content ourselves with using any means to identify the 3D configuration of the native protein structure. Most methods for protein structure prediction attempt to leverage the growing database of experimentally determined structures. Having observed the known native structures of these protein sequences, we hope to abstract principles of protein sequence and structure which may be used to accurately predict the structure of novel sequences. This process draws heavily on data-analytic techniques and statistical reasoning.

A more careful discussion of this branch of study, and indeed a precise formulation of the problem, will be given in Chapter 3. However, we may summarize by saying that the goal of accurate predictions for arbitrary sequences remains elusive. In the absence of absolute accuracy, it is therefore equally important that methods reliably indicate this inaccuracy, providing a degree of confidence which may be ascribed to their predictions. In other words, methods are desired which indicate predictions or portions of predictions that can be taken as credible, and those which must be viewed with skepticism. Once again, the techniques of statistical inference provide a solution.

Even for a hypothetical prediction algorithm with complete accuracy, we might impose further *desideratum*. While the prediction methodology may not reflect the physical process of folding, it should be interpretable, so that prediction may be used to gain further insight into the important factors of protein folding. This criteria will continue to grow in importance as large numbers of protein folds are determined

experimentally (Burley et al., 1999; Montelione and Anderson, 1999), and protein structure tools are applied to aid in understanding folding and function, and even designing new proteins, rather than simply making predictions.

1.3 Statement of hypothesis

The hypothesis of the research in this dissertation is that:
Bayesian inference provides a general framework for the prediction of protein structure from sequence. Moreover, this framework can be practically realized via the development of two components:

(i) Statistical models of sequence/structure relationships estimated from databases of experimentally determined protein structures

(ii) Computational methods, especially Monte Carlo methods, for inference with these models.

In combination with these components, Bayesian inference provides a set of tools for accurate predictions of aspects of protein structure from sequence, including reliable estimates of prediction uncertainty.
In order to demonstrate and evaluate this hypothesis, this dissertation contains the following:

1) A general framework for protein structure prediction problems via probabilistic modeling and Bayesian inference.

2) A set of parameterized sequence/structure models which can be estimated from experimental protein structure databases.

3) Computational tools for practical implementation of Bayesian inference using these models.

4) Evaluation of this approach on two different formulations of protein structure prediction: prediction of secondary structure, and prediction of β-sheet topology and tertiary contacts.

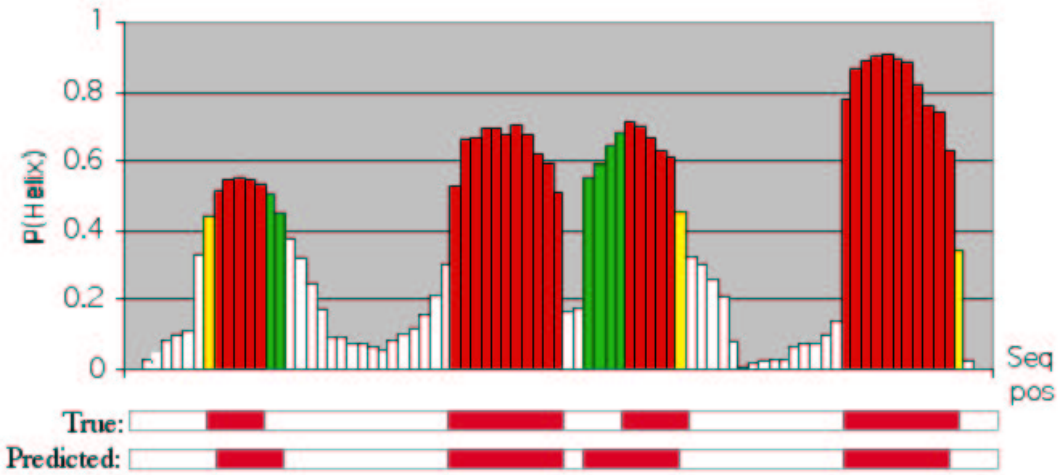


Figure 1.2: Secondary structure prediction for Cytochrome C obtained from the Bayesian segmentation algorithm. Bars indicate predicted probability of α-helical structure. Residues which take on helical conformation in the X-ray crystallography structure are shown in red. Probabilities are calculated using methods developed in Chapters 4 and 7.

1.3.1 Example

Figures 1.2 and 1.3 demonstrate an example use of the methods developed in this dissertation. A newly sequenced protein is analyzed to provide prediction of secondary structure and β-sheet topology. In Figure 1.2 the approach developed in Chapter 4 is applied to produce estimated secondary structure probabilities at each position of the sequence. In Figure 1.3, methods described in Chapter 5 produce an estimated probability-of-contact map which predicts potential tertiary interaction of β-strands forming β-sheets. As shown in Chapter 9, these outputs provide accurate and informative predictions of important features of the protein conformation based only on sequence information.

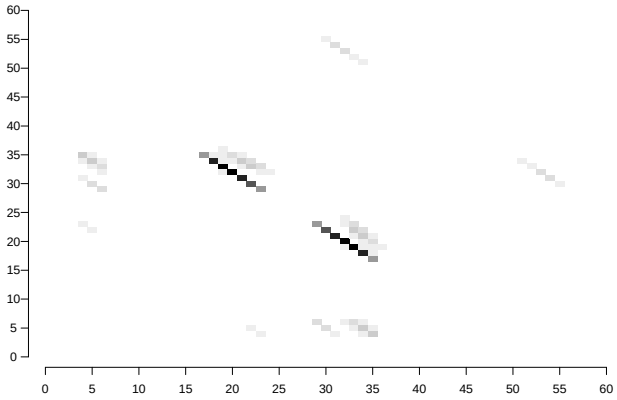


Figure 1.3: Prediction of tertiary β-sheet contact map for pancreatic trypsin inhibitor. Axes represent position in sequence, and shading of pixels (x,y) is proportional to the predicted probability of residues x,y forming contacts within a β-sheet. Probabilities are calculated using methods developed in Chapters 5 and 8.

1.4 Organization of this document

The remainder of this document is organized as follows. Chapter 2 presents a brief overview of the current understanding of protein structure and protein folding, emphasizing factors important for protein structure prediction used later in the dissertation. Chapter 3 reviews previous work in the field of protein structure prediction, with a detailed treatment of the two main problems addressed in this dissertation: prediction of the secondary structure and β-sheet topology of a protein from its amino acid sequence. Chapter 4 introduces the general framework developed in this dissertation, describing a probabilistic formulation of the protein structure prediction problem and an explicit set of probabilistic models for protein sequence/structure relationships. The application of these models for prediction of secondary structure is described. Chapter 5 generalizes the framework described in Chapter 4 to include models for *non-local* interactions in protein sequences, and demonstrates this approach by developing models for the prediction of β-sheet topology from sequence. Chapter 6 diverges from the theme of protein structure prediction to formalize in an abstract way the statistical developments of this dissertation, describing new classes of stochastic models for sequences of random variables with complex dependency structure. Chapters 7 and 8 provide the algorithms required to utilize the models of

Chapters 4-6 in practical applications. Chapter 9 discusses the evaluation of these models in support of the dissertation thesis given in Section 1.3, and Chapter 10 provides some concluding remarks as well as a perspective on future work.

This dissertation crosses several disciplines and is expected to make contributions to the fields of bioinformatics and computational molecular biology, statistics, and protein science. Chapter 2 provides a review of basic ideas of protein structure and folding, which may be useful to statisticians and computer scientists lacking formal exposure to molecular biology. Chapters 4-8 form the core of this dissertation. Bioinformaticians interested primarily in protein structure prediction may safely skip Chapter 6 and much of Chapters 7 and 8. Statisticians interested primarily in the methodology developed here may wish to skip directly to Chapters 6-8. Throughout, a working knowledge of probability theory at the undergraduate level is assumed.

Chapter 2

Protein Structure and Protein Folding

This dissertation develops a statistical methodology for addressing the problem of protein structure prediction. In this Chapter I present a brief introduction to the basic concepts of protein structure and review our current understanding of the biophysical factors important for protein folding. This discussion is necessarily abbreviated, and more thorough discussions can be found in (Stryer, 1995; Branden and Tooze, 1999; Creighton, 1993) which are still relatively accessible to non-experts.

2.1 Basics of protein primary, secondary, and tertiary structure

A protein sequence is a linear hetero-polymer, meaning simply that it is an unbranched chain of molecules where each "link" in the chain is one of the twenty *amino acids* (see Figure 2.1). The amino acids in a sequence are linked by asymmetric peptide bonds, allowing the designation of a beginning (N-terminus) and end (C-terminus) to the chain. Because each amino acid can be denoted by a canonical letter of the alphabet, this sequence of molecules can be represented succinctly by a

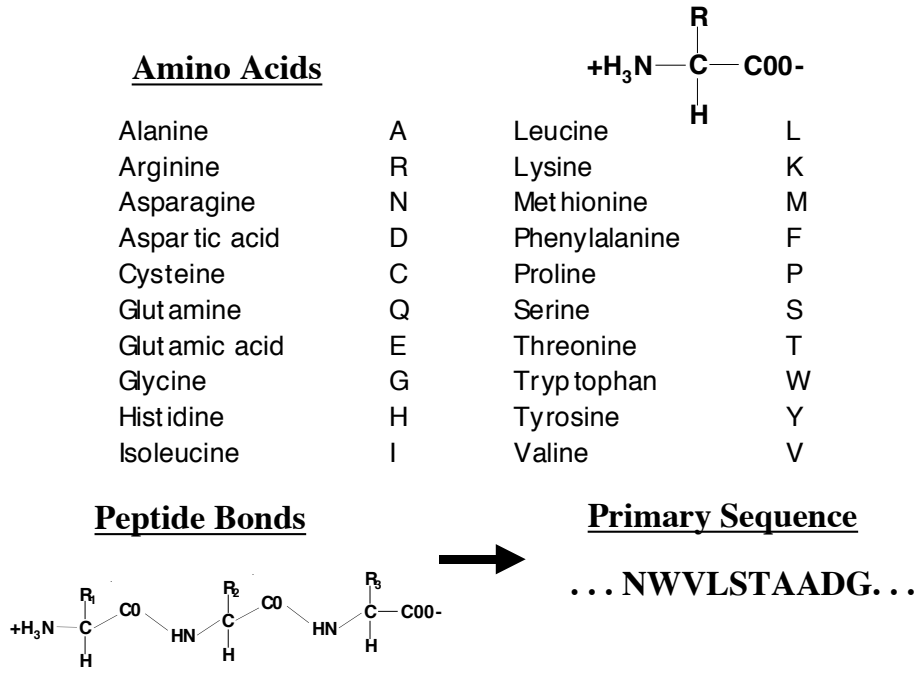


Figure 2.1: The basic components of protein structure. Proteins are made up of twenty naturally occurring amino acids linked by peptide bonds to form linear polymers. Each amino acid is represented by a letter of the alphabet to produce a protein sequence.

sequence of letters (Figure 2.1). This sequence, along with the linear chain of peptide-bonded amino acids it represents, is called the *primary* structure of the protein. In its native environment, a protein sequence folds into a compact structure in 3 dimensions (Figure 2.2), the *tertiary* structure. It is a general property of proteins that the primary sequence uniquely specifies the folded tertiary structure (Anfinsen, 1973), although examples exist where other factors may be required for efficient folding. It is this structure which provides the scaffolding for chemical or structural activity of the protein *in vivo*. Within the tertiary structure, regular conformations of the C_α backbone are observed. Figure 2.3 shows the two most commonly occurring conformations, a helical conformation known as an *α-helix* and an extended conformation known as a *β-strand*. As shown in Figure 2.3, β-strands are joined together by hydrogen bonds to form *β-sheets*. Together, such local conformations are referred to as the *secondary* structure of the protein. Protein structure can be viewed hierarchically

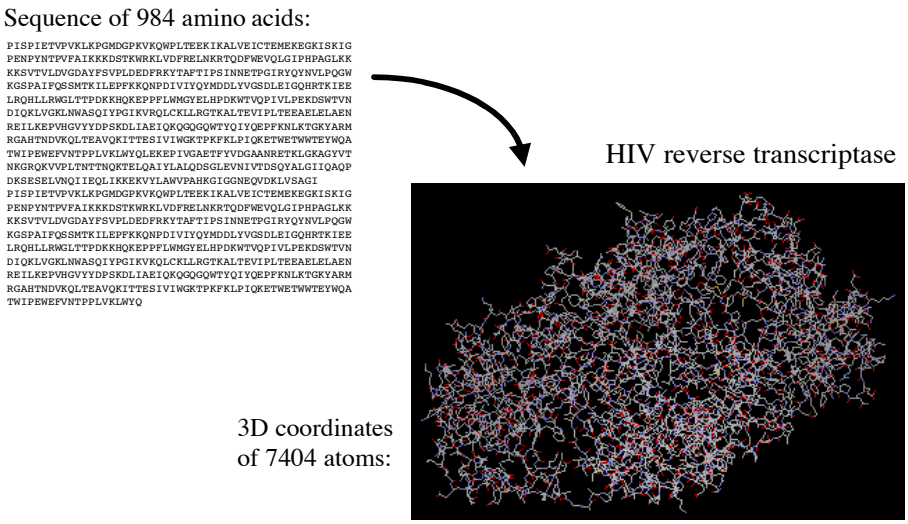
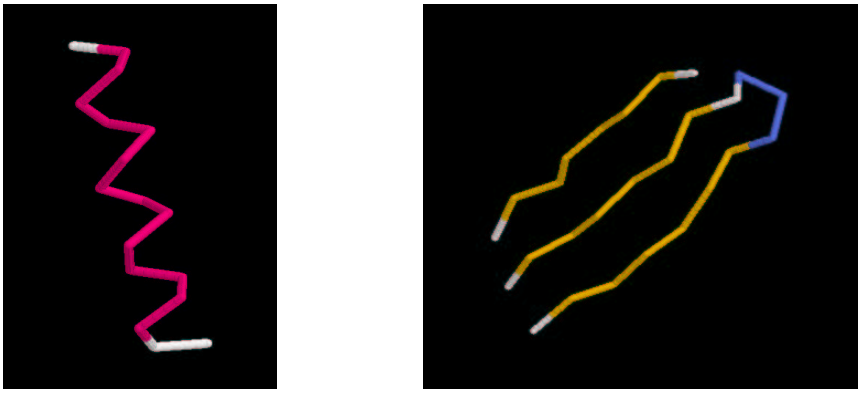


Figure 2.2: Protein folding. An amino acid sequence folds into a unique compact structure in 3 dimensions. The example protein shown is HIV reverse transcriptase, a DNA polymerase required for HIV replication and therefore a target for pharmaceutical development.



(a) (b)

Figure 2.3: Protein secondary structure. (a) *α-helix* - The B helix from sperm whale myoglobin (5mbn). (b) *β-sheet* - Three β-strands of the D β-sheet in mouse immunoglobulin (1a6w). Each β-strand forms hydrogen bonds with neighboring strands to form the sheet. The β-strands are antiparallel to each other, with two forming a β-hairpin and one coming from a sequentially distant region of the sequence. β-strands may also form parallel β-sheets (not shown).

as elements of secondary structure (α-helices and β-sheets) packed together to form a tertiary fold. Such folds may be classified based on a number of criteria including secondary structure content (all-α, all-β, $\alpha+\beta$, α/β, etc.) and the pattern into which these secondary structure elements pack, known as *super-secondary* structure. Examples of super-secondary structure are coiled-coils, which involve multiple α-helices entwined in a super-helical structure, and α-β barrels, made up of concentric rings of β-sheets and α-helices. Such classifications have become an important tool for studying evolutionary and functional relationships between large numbers of protein folds (Murzin et al., 1995).

2.2 Protein folding

The process by which a linear protein sequence folds into its final tertiary structure is known as protein *folding* (Figure 2.2). Protein folding is a subject of intense research aimed at understanding the underlying biophysical mechanisms which drive it. Forces involved in protein folding can be roughly divided into *local* effects and *non-local* effects.

2.2.1 Local effects

Amino acid propensities

Local forces in protein folding refer to the contribution made by individual amino acids of the primary sequence towards determining the conformation of the backbone at their respective positions. These forces are side chain dependent, giving rise to the widely accepted idea that amino acids have varying "propensities" for different types of secondary structure. These propensities were first observed in statistical analyses of experimental structures (Chou and Fasman, 1974b; Levitt, 1978), revealed by the different frequencies of occurrence of the various amino acids in each type of secondary structure (see Figure 2.4). In the last decade, experimental work has demonstrated that substitution of different amino acids alters thermodynamic stability of folded α-helices (Padmanabhan et al., 1990; Lyu et al., 1990) and β-sheets (Kim and Berg,

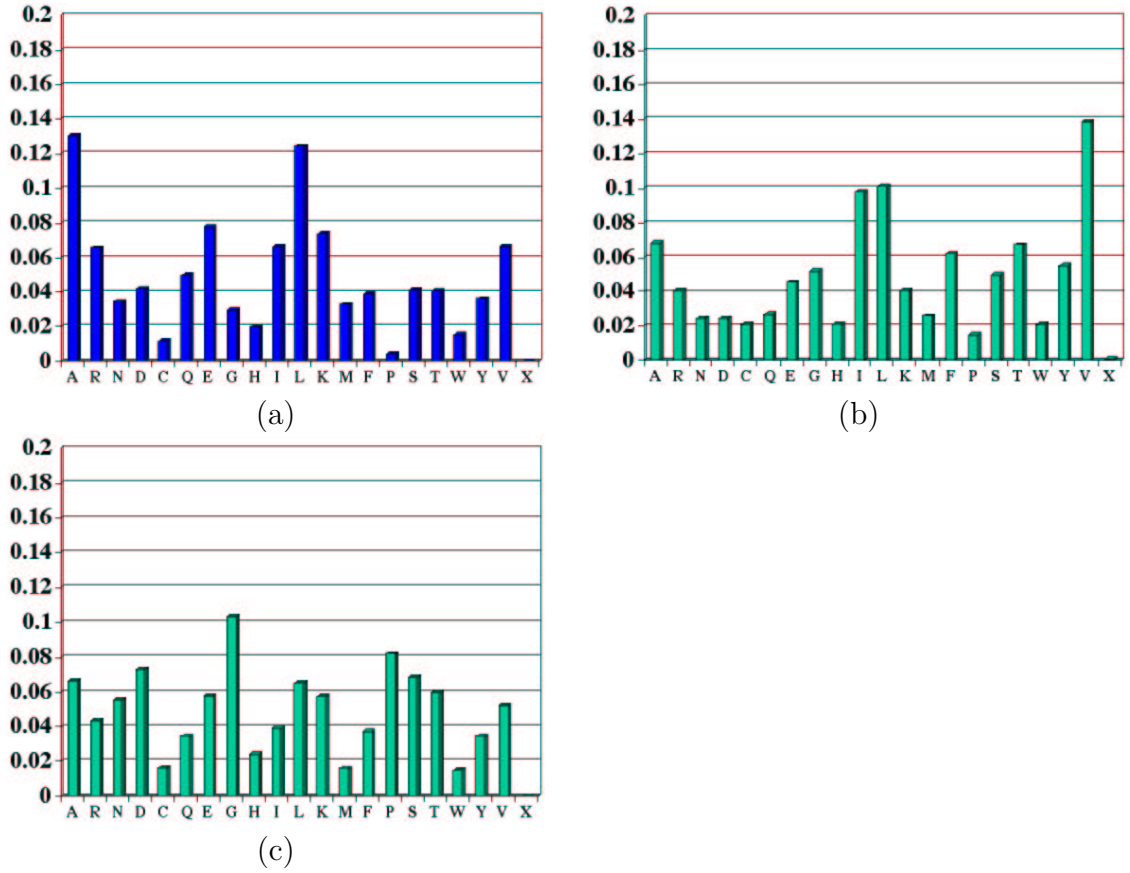


Figure 2.4: Marginal frequencies of the twenty amino acids measured at internal positions for (a) α-helices, (b) β-strands, and (c) loops/coils. Frequencies were calculated from the database of proteins described in Chapter 9.

1993; Minor and Kim, 1994b; Smith et al., 1994), using peptide and protein host-guest studies. Such experiments provide a physical meaning for the notion of propensity. These amino acid propensities, either statistical or physical, form the basis of essentially all successful methods for prediction of protein secondary structure from sequence (Chapter 3). Although secondary structure propensities have been observed experimentally as changes in free energy differences ($\Delta\Delta G$), this does not illuminate the physical basis for these changes. The exact mechanism by which different side chains yield different contributions to the free energy of secondary structure formation remains somewhat uncertain. For some side chains such as Proline, the mechanism

is clear: the Proline side chain forms a covalent bond with its amino nitrogen, preventing formation of hydrogen bonds and constraining backbone rotation. This effect makes Proline unfavorable in all but the first 4 positions of an α-helix (Kim and Kang, 1999). Glycine, which has no side chain, can be expected to disrupt regular secondary structure based on entropic arguments. Glycine also provides little steric hindrance and hence is common in positions requiring sharp changes in backbone conformation such as β-turns (Wilmot and Thornton, 1988; Hutchinson and Thornton, 1994).

The physicochemical basis for the variations observed in other side chains is somewhat less clear. An important factor in determining α-helical propensity is likely to be side chain conformational entropy loss (Creamer and Rose, 1992; Creamer and Rose, 1994; Doig and Sternberg, 1995). For example, large or β-branched side chains in α-helices encounter steric clashes with backbone atoms in some rotamer positions. Because such rotamers remain unpopulated (McGregor et al., 1987), these residues experience less conformational entropy in an α-helical backbone conformation than in a random coil, decreasing the relative free energy. Examination of Figure 2.4 lends some support to this explanation, where Val, Ile, Phe, and Tyr have lower occurrence in helices than Leu. Similar entropic arguments may be made to explain β-strands propensities (Baldwin and Rose, 1999a; Street and Mayo, 1999).

Side chain entropy is not the only possible explanation for secondary structure propensities, however. Other proposed contributions include the amount of side chain hydrophobic surface area buried in α-helices (Blaber et al., 1993), side chain steric screening of solvent competition for main-chain hydrogen bonding in β-sheets (Bai and Englander, 1994), and side chain screening and desolvation of backbone polar atom electrostatic interactions (Avbelj and Moult, 1995; Avbelj and Fele, 1998; Baldwin and Rose, 1999a).

The extent to which each of these effects contributes to overall propensity remains unresolved (Avbelj and Fele, 1998). The difficulty in distinguishing such effects experimentally makes the question a difficult one; however, recent theoretical studies indicate that conformational entropy may be a dominant factor in both α-helices and β-strands (Creamer and Rose, 1992; Creamer and Rose, 1994; Street and Mayo, 1999). It is worth pointing out that *all* side chain entropic effects are destabilizing, and so

the relative contributions of these terms have implications for different views on the driving forces of protein folding. For example, if secondary structure propensity is in fact largely determined by entropic effects, other (backbone) forces must be driving the formation of these elements. The most likely factors are hydrogen bonding and burial of hydrophobic surface area, corresponding to the *backbone centric* and *side chain centric* views of protein folding, respectively (Dill, 1999).

Position-specific propensities

Like Proline described above, other amino acids exhibit a varying propensity for α-helix formation depending on their position within the helix. In particular, the N- and C-terminal positions of α-helices have been shown to have distinctive amino acid propensities, measured both statistically (Chou and Fasman, 1974a; Presta and Rose, 1988; Richardson and Richardson, 1988; Doig et al., 1997; Schmidler et al., 2000) and experimentally (Doig and Baldwin, 1995; Petukhov et al., 1998). These effects, and the biophysical mechanisms which lead to them, are referred to as *helix capping* (Aurora and Rose, 1998). A number of factors contribute to this position-dependence of propensities. These include the differences in side chain rotamer restrictions, solvent exposure, and steric contacts encountered in various positions of a helix (Petukhov et al., 1998) due to the lack of amino and carboxyl group hydrogen bonds at the 4 N- and C- terminal positions respectively, as well as specific side chain-backbone interactions (Presta and Rose, 1988; Harper and Rose, 1993; Zhou et al., 1994), and interactions with the helix dipole (Aqvist et al., 1991; Armstrong and Baldwin, 1993). A somewhat related position-specificity has been observed in the propensities of amino acids to form β-strands. In this case however, position is defined not with respect to the strand, but with respect to the β-sheet in which it occurs. In particular, edge strands of a sheet (those with only one hydrogen bonded neighboring strand) have been shown to have different amino acid propensities from those measured at internals strands within sheets (Minor and Kim, 1994a). Figure 2.5 shows some of the position-specific amino acid distributions calculated from the database described in Chapter 9. As previously observed, there is a prevalence of Pro at position N_1 and Glu and Asp at position N_2 in α-helices, and an abundance of Pro in position N_2 as

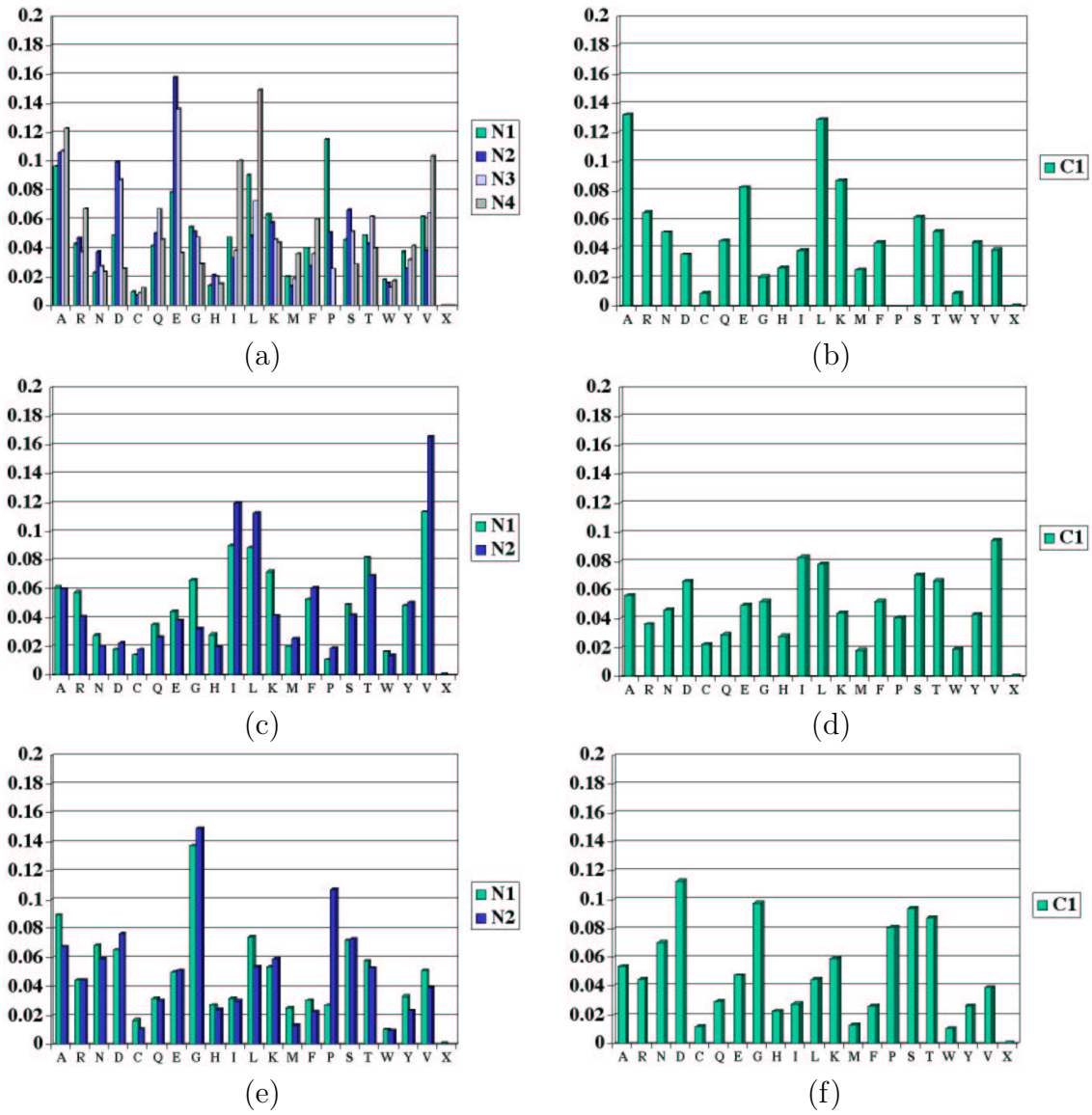


Figure 2.5: Marginal frequencies of the twenty amino acids in helical capping positions. N- and C-terminal and positions in α-helices (a,b), β-strands (c,d), and loops/coils (e,f). Frequencies were calculated from the database of proteins described in Chapter 9.

Cap position	Kullback-Leibler divergence from internal position		
	α-helix	*β-strand*	*Loop/coil*
N1	**.146**	**.053**	**.050**
N2	**.196**	.034	**.032**
N3	**.115**	.025	.020
N4	**.08**	-	.014
C4	.019	-	.008
C3	.029	.020	.012
C2	.037	.015	.011
C1	**.059**	**.077**	**.056**

Table 2.1: Kullback-Leibler divergence (cross-entropy) measured from the amino acid distribution at internal segment positions to N- and C- terminal positions. The 4 N- and C- terminal positions are shown for α-helices and loops, and 3 for β-strands (due to sparse data). Positions shown in boldface are included in the capping models described in Chapter 4. Kullback-Leibler divergence between two probability distributions $\mathbf{p}$ and $\mathbf{q}$ is defined as $KL(\mathbf{p},\mathbf{q}) = \sum_i p_i \log(\frac{p_i}{q_i})$. Data comes from $Dataset\#1$ described in Chapter 9.

an initial helix-terminating position in loops/coils[1]. Table 2.1 taken from (Schmidler et al., 2000) shows the statistical deviance among the distributions at these positions. Helix-capping effects can be seen to be among the strongest.

2.2.2 Non-local effects

Amino acid propensities are not the only forces which drive protein structure formation. A major source of difficulty in predicting protein 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

[1]The occurrence of Pro in N_2 rather than N_1 is an artifact of the structure assignments provided by DSSP (Kabsch and Sander, 1983) as described in Chapter 9. DSSP helix boundaries do not include the first and last hydrogen bonded residues in a helix, and hence the N_{cap} and C_{cap} positions are included in the loops preceding and following the α-helix, respectively.

the tertiary structure, as the sequence folds back on itself in three dimensions (Figure 2.2). Interactions between such residues are non-local from the standpoint of sequence. The relative importance of local vs. non-local effects in determining protein folds is still a subject of some debate (Baldwin and Rose, 1999a; Baldwin and Rose, 1999b; Dill, 1999). However, it is clear that non-local effects can be important. For example, (Kabsch and Sander, 1984; Cohen et al., 1993) have located identical penta- and hexa-peptides which take on very different local conformations in different proteins. Moreover, (Minor and Kim, 1996) have designed an 11 amino acid "chameleon" sequence 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 in the same protein. A possible explanation for such observations is the effect of non-local contacts in determining local structure.

A general type of non-local interaction which has been studied extensively is hydrophobic interaction (Kellis et al., 1988; Sandberg and Terwilliger, 1989; Dill, 1990). Amino acids distant in the primary sequence may form contacts in the tertiary structure if doing so helps to bury their hydrophobic surfaces against each other. Strong evidence suggests that burial of hydrophobic residues is an important factor driving protein folding (Dill, 1990; Dill, 1999). In the extreme view, hydrophobic contacts may dominate tertiary *and* secondary structure formation, and it is widely accepted that they are at least responsible for determining the packing of secondary structure in most folds.

α-helices

In addition to the general hydrophobic interactions described above, specific side chain interactions have been observed in α-helices, particularly at the $(i, i+4)$ positions brought into spatial proximity by the 3.6 residue/turn rate of the helical rotation. These interactions have been shown to contribute significantly to stabilization of α-helices both experimentally (Padmanabhan and Baldwin, 1994; Huyghues-Despointes et al., 1995; Baldwin and Rose, 1999a) and theoretically (Creamer and Rose, 1995; Shalongo and Stellwagen, 1995; Stapley et al., 1995). Stabilizing side chain interactions have also been discovered statistically by looking at pairs of co-occurring amino

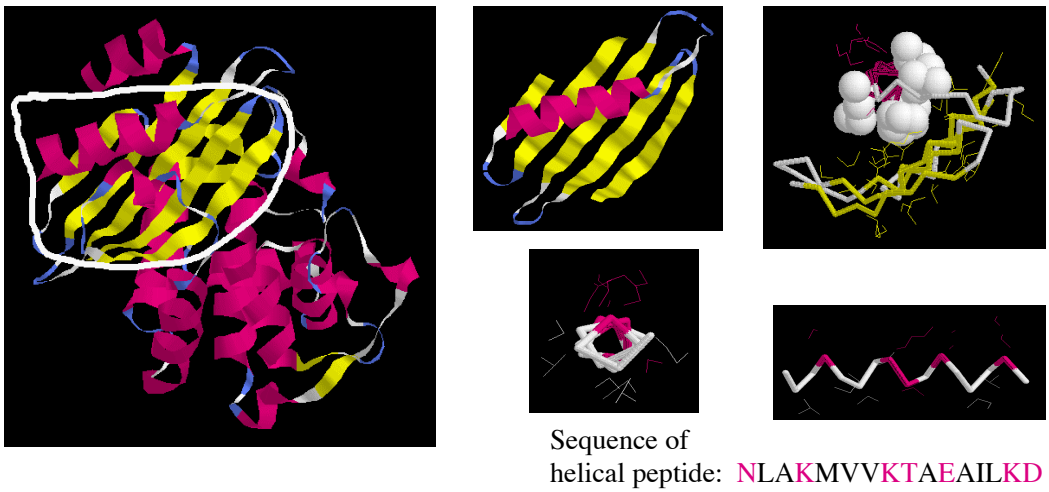
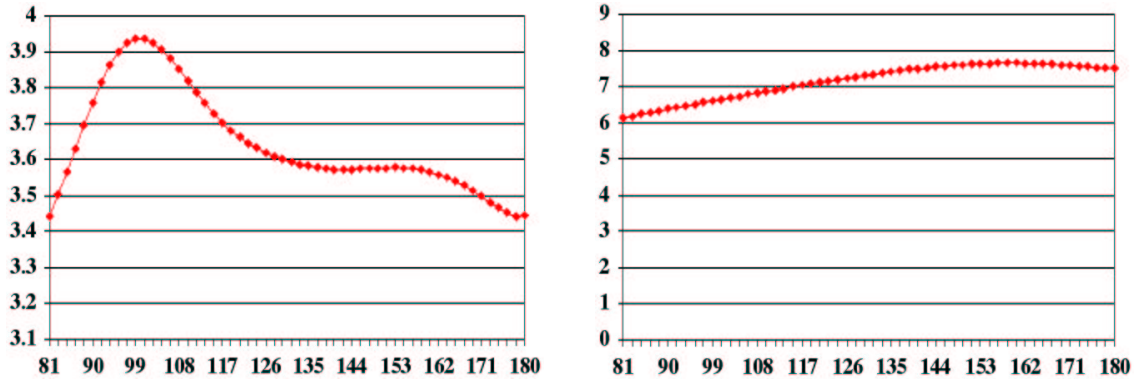


Figure 2.6: Amphipathic α-helix from β-lactamase (4blm). The helix is shown in its native environment, where one side is buried against a β-sheet. Hydrophobic amino acid side chains on the helix are shown in white, and tend to cluster on the buried side of the helix where they are not exposed to solvent. When the helix is viewed alone, these hydrophobic amino acids are seen to induce an approximate periodicity in the linear sequence, giving clues about the underlying backbone conformation.

acids in experimental structures (Klingler and Brutlag, 1994), findings later verified experimentally. General hydrophobic interactions also lead to an interesting set of side chain patterns within α-helices. Because of the periodic nature of the helical backbone, amino acids at (for example) positions $(i, i+4)$ share a common chemical environment. This common environment, in combination with possible hydrophobic side chain interactions (Creamer and Rose, 1995), leads to correlations in hydrophobicity at positions along a helix. This effect is most striking in amphipathic α-helices, where one side is buried and the other exposed to solvent (Figure 2.6). Such periodic correlations have been quantified using Fourier transforms, yielding the *hydrophobic moment* of helical subsequences (Eisenberg et al., 1982; Eisenberg et al., 1984a; Eisenberg et al., 1984b). Figure 2.7 shows a hydrophobic moment plot for α-helices extracted from the dataset described in Chapter 9. The hydrophobic moment is calculated as the modulus of the Fourier transform, and can also be given a geometric interpretation (Eisenberg et al., 1982). The quantification of amino acid hydrophobicity required for such analyses is somewhat subjective however, as a large number



(a) (b)

Figure 2.7: Hydrophobic moment plots for (a) α-helices and (b) β-strands. The hydrophobic moment (Eisenberg et al., 1982) is defined as the modulus of the Fourier transform: $\mu(\delta) = \{[\sum_{n=1}^{N} H_n \sin(\delta n)]^2 + [\sum_{n=1}^{N} H_n \cos(\delta n)]^2\}^{\frac{1}{2}}$ Peaks are observable at approximately $100°$ for α-helices and $150-160°$ for β-strands, corresponding to the backbone rotations of 3.6 residues/turn and 2.1 residues/turn, respectively. Helices and strands were extracted from the dataset described in Chapter 9.

of experimental, theoretical, and empirical scales have been reported - (Cornett et al., 1987) cite 38 such scales from the literature.

These general and specific side chain interactions lead to an observed increase in the frequency of certain pairs of amino acids at appropriately spaced positions along the helical backbone as seen in (Klingler and Brutlag, 1994) and Figure 2.8, when measured relative to the marginal propensities shown in Figure 2.4. It is difficult to distinguish which correlations may be explained by actual side chain interactions and which can simply be ascribed to similarity of chemical environment of the type indicated by Figures 2.6 and 2.7. Entropic arguments suggest that some types of side chain interactions may be rare (Padmanabhan and Baldwin, 1994). However, for purposes of predicting α-helices from sequence the distinction is irrelevant (Chapter 4).

The side chain and hydrophobic interactions in α-helices just described are non-local in the sense that they depend on amino acids which are not immediately adjacent in the primary sequence. However, these interactions are still "semi"-local, because they occur within the same secondary structural segment. A second type of side chain

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	1.342	1.09	0.751	0.609	0.954	0.909	0.762	0.982	0.998	0.996	1.004	1.083	1.222	0.878	0.257	1.014	0.846	1.126	0.96	1.001
R	0.923	0.972	1.399	1.035	0.624	1.081	1.537	0.866	0.84	0.881	0.889	0.824	0.865	0.61	0.211	0.928	0.847	0.739	0.933	0.749
N	0.976	0.909	1.273	1.096	0.472	1.196	0.978	1.203	1.265	0.633	0.784	1.059	0.988	0.99	0.326	1.295	1.068	0.794	1.026	0.891
D	1.305	1.754	1.357	1.129	0.699	1.327	1.091	0.839	1.192	0.777	0.708	1.734	1.023	0.735	0.254	1.392	1.088	0.93	1.125	0.664
C	0.933	0.802	0.393	0.699	3.159	0.375	0.678	0.852	1.213	1.156	1.2	0.867	1.202	1.33	0.287	0.824	1.08	1.227	0.663	1.084
Q	1.042	1.13	1.457	1.86	0.643	1.217	1.189	0.786	0.98	0.686	0.861	1.181	0.795	1.055	0.254	1.077	1.27	0.58	0.91	0.644
E	1.04	1.662	1.417	0.777	0.616	1.033	1.044	0.658	1.094	0.852	0.901	1.834	0.915	0.852	0.201	1	0.929	1.111	0.832	0.614
G	1.832	0.994	0.953	1.113	1.162	1.144	0.658	1.531	1.074	1.12	1.305	0.942	1.327	1.06	0.137	1.275	1.129	0.894	1.034	1.249
H	0.662	0.865	1.091	0.626	1.213	0.772	0.94	0.773	2.093	1.134	1.122	0.634	1.231	1.143	0.159	0.914	0.932	1.069	1.226	1.156
I	0.865	1.035	1.209	0.364	1.734	0.765	0.711	0.539	1.159	1.708	1.373	0.916	1.506	1.253	0.105	0.603	1.043	0.962	1.308	1.259
L	0.874	0.906	0.693	0.505	1.294	0.809	0.594	0.705	0.9	1.236	1.411	0.767	1.057	1.21	0.097	0.746	0.819	0.933	1.121	1.082
K	0.805	0.874	0.918	1.908	0.512	1.138	1.584	0.703	0.634	0.67	0.702	1.105	0.63	0.507	0.326	0.782	0.915	0.597	0.777	0.541
M	0.941	0.95	0.898	0.532	1.757	0.795	0.692	0.973	0.821	1.371	1.415	0.825	1.443	1.189	0	0.829	0.685	0.667	1.178	0.984
F	0.866	0.878	0.904	0.515	1.795	0.601	0.354	0.678	1.548	1.155	1.39	0.916	1.645	1.346	0.039	0.869	0.69	1.151	1.008	1.46
P	2.271	1.766	1.21	1.176	1.006	1.489	1.696	1.008	0.877	1.157	1.002	2.004	1.204	1.18	0	1.253	1.278	1.451	1.395	1.233
S	1.229	1.003	1.238	0.729	0.824	1.206	0.739	1.05	1.142	0.926	0.94	1.088	0.806	0.805	0.244	1.624	1.09	0.849	0.821	0.767
T	1.05	1.034	1.612	0.796	0.84	1.098	0.845	0.899	1.065	1.208	1.284	0.983	1.256	1.116	0.071	0.828	1.052	1.255	1.092	1.061
W	0.751	0.9	1.191	0.426	0.526	1.043	0.822	0.838	0.778	0.834	1.29	0.966	1.201	1.534	0.207	1.061	1.125	1.643	1.116	1.353
Y	0.763	1.054	1.097	0.44	1.4	1.17	0.626	0.634	1.349	1.16	1.128	1.339	1.122	1.028	0.218	0.821	0.856	1.116	1.162	1.366
V	1.031	1.002	1.026	0.489	1.71	0.708	0.719	0.731	0.925	1.289	1.603	1.068	1.508	1.129	0.173	0.788	0.968	1.022	1.214	1.359

Figure 2.8: Odds ratios ($\frac{P(A_1,A_2)}{P(A_1)P(A_2)}$) for occurrence of amino acid pairs in α-helices at positions $i, i+4$. Ratios over 1.5 are highlighted. Helices were extracted from the dataset described in Chapter 9

interaction has been observed in α-helices which is truly non-local in sequence. This involves interaction between side chains in distinct α-helices which pack together to form coiled-coils. Side chains at specific positions in these neighboring helices have been shown to interact (Krylov et al., 1994), and in fact to determine specificity for the type of tertiary fold. I refer to such interactions as *inter*-segment interactions, in contrast with the *intra*-segment interactions described above. These inter-segment non-local interactions, and those discussed below for β-sheets, are difficult to model in standard protein structure prediction schemes. This issue will be taken up again in Chapter 5.

β-strands

The same type of periodicity in side chain environment discussed for α-helices also holds for β-strands, and similar hydrophobic patterns can be observed. However the differences in rate of backbone rotation between α-helices and β-strands (Figure 2.3) lead to a different observed period for β-strands. In particular, β-strands alternate by placing every second side chain on the same side of a sheet. Combined with the "twisting" of β-sheets induced by hydrogen bond angles (Creighton, 1993), this leads to a period of approximately 2.1 in β-strands (Eisenberg et al., 1984b), s can be seen

in the hydrophobic moment plot (Figure 2.7).

β-sheets

As with α-helices forming coiled-coils, β-strands forming β-sheets exhibit *inter*-segment side chain interactions in addition to the intra-segment hydrophobic periodicity. In fact, there is now substantial evidence that inter-segment side chain interactions between neighboring strands lend specificity and stability to formation of β-sheets during protein folding. This evidence is provided by both statistical (Lifson and Sander, 1980; Wouters and Curmi, 1995; Hutchinson et al., 1998) and experimental (Smith and Regan, 1995; De Alba et al., 1997) analyses, and has recently been used advantageously in the design of novel peptides which fold into β-sheets (Kortemme et al., 1998; De Alba et al., 1999). Statistical analyses and simulation experiments also show that these side chain interactions in β-sheets are context-dependent, with different interactions being energetically favorable in hydrogen bonded and non-hydrogen bonded amino acid pairs (Wouters and Curmi, 1995; Hutchinson et al., 1998). Combined with the edge-dependence of amino acid propensities in β-sheets discussed in Section 2.2.1, this suggests that non-local factors play a particularly important role in the formation of β-sheets. Again, these types of non-local, inter-segment interactions are difficult to incorporate into standard protein structure prediction schemes. We will return to this issue in Section 3.2 of Chapter 3, and again in Chapter 5.

Disulfide bonds

A final set of non-local interactions worth mentioning briefly are those involving disulfide bonds. In contrast to the other interactions discussed, disulfide bonds are covalent bonds formed between two Cysteine residues to form a Cystine. These residues are often quite distant in sequence, and the Cys-Cys bond that forms when the two residues are in close proximity in the tertiary structure can significantly stabilize the folded protein. Disulfide bonds commonly occur in secreted (extracellular) proteins (Branden and Tooze, 1999; Stryer, 1995).

Chapter 3

Previous Work in Protein Structure Prediction

Chapter 2 discussed the basic principles of the structure of proteins. In this Chapter I review the problem of *predicting* this structure from the amino acid sequence, the central problem addressed in this dissertation. The field of protein structure prediction has a long history and enormous literature, and the discussion here is necessarily limited. Emphasis will be on specific formulations of the problem considered later in the dissertation. The interested reader is referred to (Sternberg, 1996) for a broader and more comprehensive introduction to the field.

Protein structure prediction methods can be broadly divided into two classes, physical and statistical. On the one hand are methods which attempt to model the physical determinants of protein structure at the molecular level. The most extreme examples of the physical approach are attempts to predict protein structure by molecular dynamics simulation (Leach, 1996; Frenkel and Smit, 1996) of the physical protein folding process (Levitt and Sharon, 1988; Duan and Kollman, 1998). Currently, computational considerations make such approaches impractical for *ab initio* protein structure prediction, although promising progress is being made (Duan and Kollman, 1998; Doniach and Eastman, 1999). I do not attempt to provide an overview here of molecular dynamics simulation, as the literature is vast and the approaches developed in this dissertation are quite different.

The second class of methods for protein structure prediction are statistical in nature. The common thread shared by these approaches is an attempt to leverage the existing database of experimentally determined structures (Bernstein et al., 1977) in order to infer structure for new sequences. Statistical methods thus broadly encompass techniques such as sequence homology search followed by structural homology modeling, threading and fold recognition, secondary structure prediction, and a variety of other methods which at some level involve fitting statistical models to database structures.

These two classes of methods, physical and statistical, are of course not mutually exclusive. The most successful approaches to date rely on a combination of physicochemical principles and experimental structure data in order to develop predictive models. Using the understanding of protein folding principles developed in Chapter 2 to provide the general structure of predictive models, and then estimating the model parameters from experimental data, is a central theme of the work presented in this dissertation.

The methodology developed in this dissertation will be demonstrated using two particular problems in protein structure prediction, which represent intermediate steps along the path to full tertiary structure prediction. These two problems are (*i*) the prediction of secondary structure from sequence, and (*ii*) the prediction of β-sheet tertiary contacts and topology from sequence. In this Chapter I provide a review of the history and current status of work in these two areas. The Chapter concludes with a brief discussion of empirical potential functions for protein structure prediction, a topic which will serve as a basis for later discussion in Chapters 4 and 5.

3.1 Secondary structure prediction

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 3.1. This 1-dimensional summary is typically formulated as a 3-state problem, with all positions

α-helix and anti-parallel β−sheet:

Residue Sequence: NWVLSTAADMQGVVTDGMASFLDKD

Secondary Structure: LLEEEELLLLHHHHHHHHHHLHHHL

Figure 3.1: The secondary structure of a protein sequence. Secondary structure of a protein is defined by the local backbone conformation at each position. Secondary structure elements of greatest interest include α-helices (shown in red) and extended β-strands which come together to form β-sheets (shown in yellow). These are represented as H and E respectively in the 1-dimensional summary.

classified as being in either α-helix (designated by H), extended β-strand (E), or loop/coil (L) conformation. Hence the problem can be viewed abstractly as a mapping from a 20^n-dimensional amino acid sequence space to a 3^n-dimensional structural sequence space, for a sequence of length n. Accurate secondary structure predictions are of considerable interest, in part because knowledge of the location of secondary structure elements can be used for approximate folding algorithms (Monge et al., 1994; Dandekar and Argos, 1996; Friesner and Gunn, 1996; Ortiz et al., 1998; Eyrich et al., 1999b; Eyrich et al., 1999a; Zhu and Braun, 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. I discuss several important historical and modern approaches here (see (Barton, 1995; King and Sternberg, 1996) for additional references.)

3.1.1 Early methods

Essentially all methods for predicting protein secondary structure from sequence are based on the observation that different amino acids have different *propensities* for each type of secondary structure, as described in Section 2.2.1. As discussed, such propensities can be understood both as physical effects measured as free energy contributions to thermodynamic stability, and statistical effects measured as marginal distributions of amino acids conditional on secondary structure. Statistical estimates of propensities are most commonly used in the prediction schemes discussed here, although attempts have been made to base predictions on experimentally determined parameters (Qian, 1996) as well.

Chou-Fasman

Among the earliest published methods to utilize these amino acid propensities for secondary structure prediction were the Chou-Fasman (Chou and Fasman, 1974b) and Garnier-Osguthorpe-Robson (GOR) (Garnier et al., 1978) approaches. (Chou and Fasman, 1974a) defined a conformational parameter for each amino acid based on the observed frequency of occurrence in each secondary structural type. These parameters were motivated by work on helix-coil transition theories, but were simply empirical frequencies defined as:

$$P_s(a) = \frac{n_{a,s}/n_{.,s}}{n_{a,.}/n_{.,.}} \tag{3.1}$$

where a is an amino acid, $s \in \{H, E, L\}$ a class of secondary structure, $n_{x,y}$ the empirical count of amino acids of type x seen in secondary structure y, and $n_{x,.}$ represents the marginalization $\sum_y n_{x,y}$. This is seen to be an estimate of the likelihood ratio $\frac{P(a|s)}{P(a)}$, or alternatively of the posterior distribution $P(s \mid a)$ assuming a uniform prior over s. Multiplication of these parameters for a series of neighboring amino acids is seen to give a simple sequence likelihood calculated under the assumption of conditional independence. Based on these estimated conformational parameters, (Chou and Fasman, 1974b) divided the twenty amino acids into distinct classes shown

Symbol	Class	Amino Acids
H_α	strong α-former	EAL
h_α	α-former	HMQWVF
I_α	weak α-former	KI
i_α	α-indifferent	DTSRC
b_α	α-breaker	NY
B_α	strong α-breaker	PG
H_β	strong β-former	MVI
h_β	β-former	CYFQLTW
I_β	weak β-former	A
i_β	β-indifferent	RGD
b_β	β-breaker	KSHNP
B_β	strong β-breaker	E

Table 3.1: Assignment of amino acids to secondary structure classes used by the Chou-Fasman algorithm (Chou and Fasman, 1974b).

in Table 3.1, such as "strong α-helix former", "weak β-strand former", and "α-helix breaker". These classifications can be seen to roughly conform to the modern understanding of physical amino acid propensities discussed in Section 2.2.1. Based on this discrete classification of the amino acids, (Chou and Fasman, 1974b) scores were assigned to tetra- and hexa-peptide subsequences in a new protein sequence based on the number of amino acids of each class contained in the peptide. High scoring regions were considered "nucleation sites" for potential α-helices and β-strands, and a set of heuristic rules were defined for extension of nucleation sites in each direction, termination criteria, and resolution of overlaps. These rules provided one of the first algorithms for prediction of 3-state secondary structure from sequence. (Chou and Fasman, 1974b) reported accuracies of 77-80% on a per-residue basis (measurement of predictive accuracy will be discussed in detail in Chapter 9). This was later shown to be a significant overestimate by (Nishikawa, 1983), whose experiments estimated the accuracy at about 55%.

Garnier-Osguthorpe-Robson

A second early method for secondary structure prediction, also based on observed amino acid frequencies, was developed by (Garnier et al., 1978). Described as an information-theoretic approach, the Garnier-Osguthorpe-Robson (GOR) method defines the "information content" of an amino acid a in a particular secondary structure s as $I(s;a) = \log \frac{P(s|a)}{P(s)}$, which is easily seen to be equivalent to (the log of) the Chou-Fasman parameter. (Garnier et al., 1978) go on to define the "information difference" $I(\Delta s;a) = I(s,a) - I(\neg s;a)$, which can be seen to be the likelihood ratio $\log \frac{P(r|S)}{P(r|\neg s)}$. Like the Chou-Fasman parameter, this ratio can be evaluated on a residue by residue basis for predicting a new sequence. (Garnier et al., 1978) applied this for a local "window" around each residue, predicting the residue based on it's local subsequence. This was done by summing the information differences for each position in the window, resulting in an implicit assumption of conditional independence in the likelihood. Higher order dependencies have since been introduced into newer versions of the GOR algorithm (Garnier et al., 1996). Predictions for the entire sequence were then obtained by "sliding" this window along the length of the sequence. As with the method of (Chou and Fasman, 1974b), accuracies reported by (Garnier et al., 1978) were shown to be overestimates by (Nishikawa, 1983), who estimated the accuracy of the original GOR method at about 55%.

3.1.2 Modern methods

Perhaps one of the most influential contributions of the GOR method was the notion of window-based prediction. By reducing the problem to one of mapping a fixed-length input vector to a set of output classes, (Garnier et al., 1978) opened the door to application of a variety of statistical classification techniques developed in the statistics, machine learning, and pattern recognition communities. A wide range of such techniques has been applied over the years (Qian and Sejnowski, 1988; Holley and Karplus, 1989; Stolorz et al., 1992; Munson et al., 1993a; Munson et al., 1993b; Munson et al., 1994; Rost and Sander, 1993b; Yi and Lander, 1993; Mehta et al., 1995; Salamov and Solovyev, 1995; King and Sternberg, 1996; Riis and Krogh,

1996; Chandonia and Karplus, 1999). Many of the more successful approaches have been based on non-linear predictive models such as neural networks and nearest-neighbor classifiers. A recognized advantage of these approaches was the ability to capture dependency among the sequence positions in an efficiently parameterized way, in contrast to the implicit conditional independence assumptions of the early methods. As discussed in Section 2.2.2, such dependencies can be an important clue towards identifying backbone conformation. As with the GOR approach, most of these window-based classifiers predict each sequence position independently based on the local surrounding subsequence. In many cases, post-prediction filtering is applied to smooth the predictions in order to remove very short predicted segments (Holley and Karplus, 1989; Rost and Sander, 1993b; Zimmermann, 1994; Frishman and Argos, 1996). I describe a few of these window-based approaches here, focusing on those which have proven most successful.

Probably the most widely recognized progress in secondary structure prediction was achieved by the PHD system (Rost and Sander, 1993b). PHD was the first rigorously validated prediction scheme to break the "70% barrier" which had previously been hypothesized for secondary structure prediction. (Discussion of percentage accuracy measurements will be taken up in Chapter 9.) PHD was based on multiple neural network models, which were improved by the inclusion of a multiple alignment of homologous sequences (Rost and Sander, 1993a; Rost and Sander, 1993b; Rost and Sander, 1994) to the sequence of interest. However, PHD applied to single protein sequences performed no better than other existing methods, at about 63% (Rost and Sander, 1993b). The use of homologous sequences has proven to be a powerful source of information in secondary structure prediction (Section 3.1.3).

Further improvements in the accuracy of single sequence predictions were demonstrated by the use of nearest-neighbor algorithms (Yi and Lander, 1993; Frishman and Argos, 1996). (Yi and Lander, 1993) had the advantage of giving theoretically justifiable prediction probabilities as estimates of prediction confidence. (Frishman and Argos, 1996) were among the first to incorporate non-local information into an algorithm which achieved competitive performance. This was done by using pair potentials for β-sheets (see Section 3.2 below) and combining maximum pairing scores

over the rest of the sequence with a standard nearest-neighbor classifier. This incorporation of non-local information is cited as a source of the algorithm's predictive accuracy, both nearest-neighbor algorithms (Yi and Lander, 1993; Frishman and Argos, 1996) report accuracy levels of approximately 68%.

Generalizing the work on nearest-neighbor algorithms (Yi and Lander, 1993; Salamov and Solovyev, 1995; Frishman and Argos, 1996), (Salamov and Solovyev, 1997) showed that use of local sequence alignment as a technique for identifying "nearest neighbors" yielded improvements in single sequence accuracy to 71%.

3.1.3 Use of evolutionary information

Discussion to this point has centered on predictions for individual sequences. In the last decade, a number of authors have demonstrated significant increases in accuracy by the use of multiple sequence alignments in prediction (Rost and Sander, 1993a; Rost and Sander, 1994; Salamov and Solovyev, 1995; Di Francesco et al., 1996). This is achieved by replacing each amino acid in the single input sequence by a vector of amino acid frequencies computed over the alignment column at that position. Typically other indicators for insertions/deletions are included as well. Intuitively, alignments of multiple homologous sequences provide information about tolerance of the structure to mutations, insertions, and deletions. Since highly homologous sequences will represent other viable forms of the protein, the underlying structure is presumed to be essentially identical. Thus substitutions occurring in positions of the alignment can be assumed not to destabilize the secondary structure at that position significantly. The variety of allowed substitutions is therefore a valuable clue towards identifying local backbone conformation.

Interestingly, the use of multiple sequence information appears to provide relatively uniform improvement across a range of published methods, leading to accuracy improvements of approximately 4-6%.

3.1.4 Segmentation algorithms

While the vast majority of work on secondary structure prediction has taken a window-based approach as described in Section 3.1.2, there have been notable exceptions. Because the methodology developed in this dissertation involves the more general approach of *segmenting* an input sequence into structural regions, I briefly describe some related work here.

The algorithm of (Chou and Fasman, 1974b) described in Section 3.1.1 can be viewed as a segmentation-type approach, based on locating and extending potential α-helix and β-strand segments. However, initial identification of these sites is identified by window-based scanning. In a similar approach, (Solovyev and Salamov, 1994) use linear discriminant analysis to predict structure of all short segments of the sequence, and then resolve conflicting predictions for overlapping segments. In contrast with the original (Chou and Fasman, 1974b) algorithm, this approach does not assume conditional independence for segment predictions, but instead uses predictive features derived from hydrophobic moments and residue pairs. (Cohen et al., 1986; Presnell et al., 1992) use deterministic pattern-matching methods to locate turns and helices, using regular expressions to identify potential segment boundaries. Although many of these approaches are similar in spirit to the models developed in this dissertation, none achieve accuracies competitive with the window-based classifiers discussed above. The methodology developed in Chapter 4 can be viewed as a unifying framework for many of the ideas motivating this earlier work, achieving accurate prediction by developing a firm theoretical grounding of probabilistic modeling and inference to synthesize these various sources of information.

(Auger and Lawrence, 1989; Liu and Lawrence, 1996) describe preliminary approaches to protein sequence segmentation which helped motivate some of the ideas in Chapter 4. (Burge and Karlin, 1997) develop a model very similar to that given by Equation 4.1 of Chapter 4 for modeling introns and exons in DNA, which has been applied to gene parsing in eukaryotic DNA with great success; other related approaches to gene prediction include (Kulp et al., 1996). Early work on biological sequence segmentation using stochastic models dates back at least to (Churchill, 1989).

In all of these approaches, no attempt has been made to generalize these ideas to include segment interactions of the type developed in Chapters 5 and 6.

3.1.5 Modeling helical correlations

As mentioned in Section 2.2.2, correlations between amino acids in α-helices due to amphipathicity and side chain interactions are well known. To provide historical context for the development of α-helix models described in Chapter 4, I point out previous attempts to use this information in prediction. (Stultz et al., 1993) use a model for amphipathic helices in their development of hidden Markov models for specific structural families. (Solovyev and Salamov, 1994) use helical residue pair frequencies and hydrophobic moments to classify short segments by linear discriminant analysis as described above, and (Munson et al., 1994) build helical periodicity into logistic regression models. (Frishman and Argos, 1996) include residue pair information for hydrogen-bonded positions in α-helices into a nearest-neighbor metric, and (Riis and Krogh, 1996) develop structured neural networks which build these helical correlations into the topology of their networks. (Berger and Wilson, 1995; Berger et al., 1995; Berger, 1995) use residue correlation in helices to identify coiled-coils based on the well-known heptad repeat. Finally, (Klingler and Brutlag, 1994; Klingler, 1996) developed local probabilistic models for correlated residues pairs in α-helices and examined their use in structure prediction. The development of this aspect of the helix models provided in Chapter 4 was originally motivated by the work of (Klingler and Brutlag, 1994; Klingler, 1996).

3.1.6 Current status

Among the methods described above, best published results are currently at the level of 68% (Yi and Lander, 1993; Frishman and Argos, 1996) to 71% (Salamov and Solovyev, 1997) for predictions based on individual protein sequences. When multiple sequence alignments are available, these approaches have reached 73% (Salamov and Solovyev, 1997) to 75% (Frishman and Argos, 1997). Evaluation methodology for estimating these prediction accuracies will be discussed in Section 9.1. Small

variations in these accuracies must of course be treated with skepticism.

Although these non-linear, window-based statistical classifiers currently perform at the highest published accuracy levels, a widely recognized drawback of all such approaches is the lack of interpretability of model parameters. Such "black-box" predictors provide little insight into the important factors in achieving better prediction, making them far less valuable as tools for studying protein folding, mechanism, and design. The methodology developed in this dissertation achieves similar levels of secondary structure prediction accuracy (Schmidler et al., 2000) and Chapter 9, while maintaining a clear interpretation from a physicochemical standpoint. In addition, this dissertation provides a unified framework for treatment of secondary structure prediction and non-local tertiary interactions simultaneously in a rigorous fashion.

3.2 Prediction of β-sheets

A step beyond secondary structure prediction but short of predicting full tertiary structure is the prediction of β-sheet topology. Because β-sheets involve the formation of specific tertiary contacts between secondary structure elements (β-strands) which are non-local in sequence, the problem is well beyond the scope of standard secondary structure prediction.

Indeed, identification of β-sheet contacts is a significant step in identifying the packing of secondary structural elements into a tertiary fold, and drastically reduces the conformational search space for the problem of predicting tertiary structure from knowledge of secondary structure. Such reduction may prove to be critical for these hierarchical approaches to protein folding. For example, (Ortiz et al., 1998) point out the need for better predictions of tertiary contacts to enable tertiary structure prediction from secondary, while (Eyrich et al., 1999b) note in particular that large β-sheet proteins are difficult to predict from knowledge of secondary structure alone, and suggest development of methods which can consider strand pairing efficiently. (Zhu and Braun, 1999) show that if the strand segments are assumed known, correct alignments between paired strands can be used to provide approximate contact distances which can then be used for tertiary structure prediction.

While α-helices appear to form primarily from local effects and have been widely studied for some time (Chapter 2), the principles of β-sheet formation are far less well understood. This is due partly to the importance of non-local effects in forming β-sheets, and partly to the fact that β-sheets are often hydrophobic and buried in the protein core, making model systems more difficult to work with experimentally. However these same factors are making a detailed understanding of β-sheet formation a subject of increasing interest, in part because of the growing implication of β-sheet misfolding and hydrophobic aggregation in important neuro-degenerative diseases (Prusiner, 1997; Prusiner, 1998; Benzinger et al., 1998; Janek et al., 1999).

3.2.1 Previous work

In contrast to secondary structure prediction, attempts to predict the topology of β-sheets from individual sequences have been somewhat rare. The majority of attempts to model β-sheets have been directed towards improving secondary structure prediction (Frishman and Argos, 1996) or developing empirical potentials for fold recognition based on known protein architectures (Hubbard, 1994; Hubbard and Park, 1995). However several authors have considered the problem of β-sheet prediction, and I briefly review their approaches here.

Statistical analysis of cross-strand pairwise frequencies in sheets goes back some time (Lifson and Sander, 1980; Wouters and Curmi, 1995; Hutchinson et al., 1998). The first serious attempt to use these statistical correlations for prediction of β-sheets from sequence appears to be (Hubbard, 1994; Hubbard and Park, 1995). Hubbard defines a set of *pairwise potentials* (see Section 3.3) of the following form:

$$energy(i,j) = \sum_{k=j-2}^{j+2} -\log\left(\frac{P(R_i,R_k)}{P(R_i)P(R_k)}\right) \tag{3.2}$$

where $P(a,b)$ is the proportion of residues pairs of the form (a,b) and $P(a)$ are marginals. This can be viewed as the log of a likelihood ratio, or equivalently as an empirical $\Delta\Delta G$ of interaction (Section 3.3). A distinct set of potential parameters are defined for parallel vs. antiparallel sheets, hydrogen-bonded vs. non-hydrogen bonded

pairs, and N- terminal vs. C-terminal directions (so $P(a,b) \neq P(b,a)$). Hubbard tests the ability of these potentials to discriminate among registers in known β-strand pairs, and reports accuracies of approximately 85%. In a similar experiment (Zhu and Braun, 1999) obtains accuracies of only 63%.

These pairwise potentials are used in (Hubbard, 1994; Hubbard and Park, 1995) for sequence-to-structure threading (Bowie et al., 1991; Jones et al., 1992) to improve fold recognition in structures containing β-sheets. However, (Hubbard, 1994) also shows that these potentials may be used to attempt to predict β-strand contacts from sequence alone. This is done by evaluating (3.2) for each pair of amino acids in the sequence. A single test sequence is shown by way of example, but no attempt is made to evaluate this approach systematically. The example shown indicates low specificity in identifying potential pairing, and Hubbard concludes that the approach is not useful for single sequences. In the presence of a significant number of multiply aligned homologous sequences, he shows that specificity is improved somewhat. The potentials developed by (Hubbard, 1994) are similar in form to those used in Chapter 5, but Hubbard does not attempt to include these in a general secondary structure prediction scheme. Thus false positives also arise from predicting strands to pair with helical regions, further decreasing specificity. Nevertheless, the potentials derived by Hubbard and others (Lifson and Sander, 1980; Wouters and Curmi, 1995; Hutchinson et al., 1998) help motivate the models developed in Chapter 5.

(Krogh and Riis, 1996) take a somewhat different approach, by training a two-window neural network to recognize correctly paired β-strands. This can be viewed as a discriminative approach for fitting the type of empirical potential given by (3.2) above. By scanning all pairs of windows in a proteins sequence, predictions similar in spirit to those of Hubbard can be obtained. (Krogh and Riis, 1996) show predictions for two example protein sequences, exhibiting very little specificity. The overwhelming number of false positives obtained lead them to speculate that inter-strand correlations are very weak. They then go on to formulate an energy function based on a weighted combination of neural network predictions of secondary structure and these predicted β-strand contact propensities, and optimize this function using simulated annealing. This approach to include pairwise potentials into full secondary structure

prediction is quite similar in spirit to the approach developed in Chapter 5. However a rigorous underlying model is not specified and the performance reported is limited.

(Frishman and Argos, 1996; Frishman and Argos, 1997) use a similar window-scanning approach in combination with a pairwise potential of the form given by (3.2). Their primary purpose is to improve secondary structure predictions, but they show two examples of using the potential to predict β-strand contact maps. As with those above, the examples reported show little specificity and no attempt is made to evaluate the contact predictions systematically.

(Asogawa, 1997) uses a Hopfield neural network to incorporate more global constraints into predictions based on the approach of (Hubbard, 1994). These constraints involve contiguity of residue structures, and can be viewed as defining a Gibbs distribution on the space of possible interactions. This approach bears some semblance to the approach of (Krogh and Riis, 1996) and that developed in Chapter 5. As with the approach of (Krogh and Riis, 1996) however, no attempt is made to evaluate the prediction of β-strand contacts rigorously, and improvements on secondary structure prediction are reported only for 2-state prediction (β-strand vs. coil), a significantly easier problem than the standard 3-state prediction described in Section 3.1. Some improvement in specificity over the approach of (Hubbard, 1994) is reported by introducing the dependencies model of the Hopfield network.

(Zhu and Braun, 1999) show that if the strand segments are assumed known, predicted pair alignments using a potential of the form (3.2) can be used to provide approximate contact distances which can then be used for tertiary structure prediction; however, they make no attempt to predict the strands, and hence the structure, from sequence ab initio.

Finally, (Gobel et al., 1994) attempt to use residue correlations from multiple sequence alignment in protein families to predict general non-local contacts in the tertiary structure, but limited results indicate the problem to be very difficult.

3.2.2 Current status

As discussed, work on prediction of β-sheets is still in its infancy. Of the literature reviewed in the last section, almost no attempt is made to evaluate the prediction of β-sheets systematically. In many cases, the authors report that performance on test cases is insufficient to warrant more extensive evaluation. While (Krogh and Riis, 1996; Asogawa, 1997) report preliminary attempts to measure effects of adding β-strand pairwise potentials on improving secondary structure prediction accuracy, no published systematic evaluation of prediction of β-sheet contacts or topology has been undertaken.

It is worth pointing out that among the approaches surveyed, only (Krogh and Riis, 1996) and (Frishman and Argos, 1996) attempt to predict β-sheet topology in the context of simultaneous full secondary structure prediction from sequence. This simultaneous prediction is important in considering competing "hypotheses" for predicted segments, allowing strong helix predictions to reduce spurious predictions of β-strand contacts.

3.3 Empirical potentials

This section briefly reviews a final topic in protein structure prediction, that of *empirical potentials* for proteins. Discussion of this topic in the current context will prove valuable for discussions undertaken in later chapters.

The importance of energy functions in the study of protein structure derives from widespread acceptance of the *thermodynamic hypothesis* (Anfinsen, 1973) which states that the native structure of a protein is the one which minimizes Gibbs free energy of the system. Hence physical models of protein structure prediction are typically formulated as problems of global energy minimization. I will not discuss the development of physically-motivated energy functions here, which involve theoretically and experimentally fitted terms for electrostatic and Van der Waals interactions, bond vibrations, and so forth. Development and validation of such potentials is an important area of research which is quite mature (Leach, 1996; Frenkel and Smit,

1996). Instead I describe the development of *empirical* potentials, which help provide links between statistical mechanical models and the probabilistic models developed in Chapters 4 and 5.

Empirical potentials for proteins are energy functions derived from statistical analysis of protein structural databases (Sippl, 1995). An example is the energy function for β-sheet amino acid pair interactions defined by (3.2). The theoretical basis for use of energy functions such as this is provided by the following relation between the probability of a state q of a system and its energy $E(q)$:

$$P(q) = Z^{-1} \exp\left(-\frac{E(q)}{kT}\right) \tag{3.3}$$

where T is the temperature, k is Boltzmann's constant, and $Z = \int_{q \in Q} \exp\left(-\frac{E(q)}{kT}\right)$ is the normalization constant (probability) or the partition function (physics) when viewed as a function of external parameters such as T. Equation (3.3) is referred to as the *Boltzmann* or *canonical* distribution, and derives from statistical physics. It holds for a system in equilibrium contact with a heat reservoir, and follows directly from the fundamental postulate of statistical mechanics (Reif, 1965).

The implications of this relation for development of protein structure potentials can be seen by simply rewriting (3.3):

$$\Delta E(q, q') = E(q) - E(q') = -kT \log\left(\frac{P(q)}{P(q')}\right) \tag{3.4}$$

This means that by estimating probabilities of occurrence $P(q)$ by empirical frequencies $\frac{n_q}{n_.}$, we can obtain estimates of free energy differences. Since empirical frequencies can be calculated directly from observed structures, it is possible to define empirical potentials for use in protein structure simulation and prediction. Such potentials have been crucial to the success of threading and fold-recognition algorithms (Bowie et al., 1991; Jones et al., 1992). Although empirical potentials cannot be expected to provide the fine resolution of atomic-level potentials, they have a number of advantages. For example, they can be defined on arbitrary parameters, and hence can be used for reduced representation models of proteins (Wilson and Doniach, 1989; Sippl,

1995). In addition, empirical potentials implicitly incorporate aspects of folded proteins which are quite difficult to model at the atomic level, such as the effects of solvent and conformational entropy.

Use of (3.4) suffices to estimate energies for all states in the system. Note that absolute free energies are meaningless, and differences may be defined with respect to an arbitrary ground state q_0 by setting $E(q_0) = 0$. For example, the amino acid propensities in Section 2.2.1 were given both statistical interpretation based on frequency of occurrence and physical interpretation based on experimental measurements of free energy differences. It is now easy to see that the former can be viewed as an estimate of the latter. Similarly, the odds ratios for pairwise frequencies given in Section 2.2.2 can be viewed as $\Delta\Delta G$'s, or free energy differences due to interaction. Experimental measurements calculated for α-helical and β-strand propensities, N-capping preferences, and side chain interactions in α-helices and β-sheets, have been compared with empirical scales estimated from protein structure databases, and show general agreement. Measurements taken on model systems are not expected to agree precisely with empirical estimates, which represent averages over possible environments in folded proteins. Finally, it is important to point out that the use of propensities estimated from databases for the purposes of secondary structure prediction (Section 3.1) and β-strand contact prediction (Section 3.2) can also be viewed as empirical free energies. In most of these prediction algorithms however, these energies are combined in *ad hoc* ways which inhibit this interpretation. In Chapter 10, I will point out that the probability models developed in this dissertation can be viewed alternatively as empirical free energy potentials, and that the Bayesian inference framework developed in Chapters 4 and 5 provides correct combination of these energies for prediction based on statistical mechanical principles.

Chapter 4

A Bayesian Framework for Protein Structure Prediction

This Chapter presents the basic framework developed in this dissertation. The first section describes broadly the formulation of protein structure prediction as a problem of Bayesian inference. Following sections develop a class of probability models for proteins based on structural segments which instantiate this framework, develop a particular set of models which capture many of the principles described in Chapter 2, and develop computational methods for prediction of secondary structure under this class of models.

4.1 Protein structure prediction as a Bayesian inference problem

The general approach of this dissertation is to formulate protein structure prediction as a problem of Bayesian inference. By specifying a joint probability model over the space of protein sequences and structures $P(Sequence, Structure)$, we reduce the problem of structure prediction to the conceptually simple task of computing the conditional or posterior distribution $P(Structure \mid Sequence)$ and from it the desired predictive quantities. For a particular probability model and a given

protein sequence, structure prediction reduces to maximization or integration over $P(Structure \mid Sequence)$. In addition to the elegance and simplicity of Bayesian estimators, their optimality properties for related problems of estimation, prediction, and pattern recognition are well established (Berger, 1985). The probabilistic formulation has the additional advantage of being flexible and extensible, enabling consistent incorporation of other relevant information as well. For example, the primary sequence is only one type of "observation" on the underlying structure, and Bayesian inference may be applied to other experimental measures on structure (Altman, 1995), or used to synthesize evidence from multiple sources.

A Bayesian view of protein structure prediction is not new, although not widely established. Many existing statistical or probabilistic approaches to structure prediction can be viewed as instances of this framework, although not all are explicitly presented as Bayesian (Stolorz et al., 1992; Stultz et al., 1993; White et al., 1994; Altman, 1995; Berger, 1995; Klingler, 1996). Many of the secondary structure prediction methods described in Chapter 3 can also be viewed as Bayesian, yielding exact or approximate posterior inference under a set of (often implicit) modeling assumptions. However, while the general principles of Bayesian inference guide many of these approaches, the realization in terms of problem formulation and parameterization, model development, and computational issues, yield significantly different methodologies. These aspects of the Bayesian framework developed in this dissertation are introduced in the following sections.

In this dissertation I develop a Bayesian approach to protein structure prediction using a parameterization of protein sequence/structure relationships in terms of structural *segments*. I develop a Bayesian approach to the assignment of these parameter values, by defining a joint probability distribution for an amino acid sequence and its structural assignment. With such a model defined, I show how to compute the conditional or posterior probability distribution over structural assignments given a new sequence via Bayesian inference, and to predict those secondary structure assignments which maximize this posterior distribution.

In Section 4.2, I define a general class of segment-based joint probability models which lend themselves to efficient exact calculation of the posterior. Section 4.3

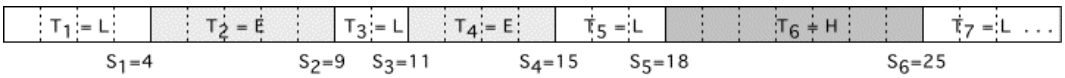


Figure 4.1: Representation of the secondary structure of a protein sequence in terms of structural segments. The parameters shown represent the segment types $T = (L, E, L, E, L, H, L, \ldots)$ and endpoints $S = (4, 9, 11, 15, 18, 25, \ldots)$. The associated structural sequence is $LLLLEEEEELLEEEELLLHHHHHHHLLL\ldots$.

provides specific models for α-helices, β-strands, and loops/coils, and shows how such models can be used to capture key aspects of protein structure formation discussed in Chapter 2. Section 4.4 describes application of these models to the prediction of protein secondary structure by extracting the relevant predictors from the posterior distribution $P(Structure \mid Sequence)$. Chapter 7 addresses computational issues and provides algorithms for computation of these predictors under the class of models developed. Evaluation of this approach based on experimental data is provided in Chapter 9. Having established the core methodology, Chapter 5 will go on to show how this framework may be generalized to incorporate non-local aspects of protein structure and move beyond secondary structure prediction into prediction of tertiary contacts.

4.2 Segment-based probability models for proteins

4.2.1 Parameterization

I begin by choosing a representation of sequence/structure relationships in proteins which is based on *segments* of secondary structure (Schmidler et al., 2000). This model is parameterized in a convenient fashion by representing the segment positions and structural types. Segment locations are denoted by the position of the last residue in the segment. The requirement that segments be contiguous implies that this parameterization uniquely identifies a set of segment locations for a given sequence. Let $R = (R_1, R_2, \ldots, R_n)$ be a sequence of n amino acid residues. and let $S = \{\ i : Struct(R_i) \neq Struct(R_{i+1})\}$ be a sequence of m positions denoting the end of each structural segment (with $S_m = n$), and $T = (T_1, T_2, \ldots, T_m)$ be the corresponding

sequence of secondary structural types[1]. Together S and T completely determine a secondary structure assignment for a given amino acid sequence. An example is given in Figure 4.1. In the case of secondary structure prediction, the quantities of interest are the values of m, $S = (S_1, S_2, \ldots, S_m)$ and $T = (T_1, T_2, \ldots, T_m)$ corresponding to the known amino acid sequence $R = (R_1, R_2, \ldots, R_n)$, that is, the number, locations and types of the secondary structural segments. The problem is to infer the values of (m, S, T) given a residue sequence R. We will refer to the set $\mathcal{S} = (m, S, T) = \{S_i, T_i\}_{i=1}^{m}$ as a *segmentation* of the sequence R.

4.2.2 Likelihood

A general class of segment-based joint distributions can now be defined over $(R, \mathcal{S})$ of the form:

$$P(R, \mathcal{S}) \propto P(\mathcal{S}) \prod_{j=1}^{m} P(R_{[S_{j-1}+1:S_j]} \mid \mathcal{S}) \tag{4.1}$$

The key aspect of this joint distribution is the decomposability of $P(R \mid \mathcal{S})$ into individual segment terms. In other words, the joint distribution may be factored by conditional independence of inter-segment residues, so that the sequence likelihood may be written as a product of segment likelihoods.

The j^{th} term in the right-hand side of (4.1) is the likelihood of the subsequence of R contained in segment j (beginning at position $S_{j-1}+1$ and ending at position S_j). The exact form of this segment likelihood is structure-dependent, and the specification of this form for each structural type amounts to developing a probabilistic *model* of the given type of segment. A particular set of models are developed below in Section 4.3. Note that this model does not assume conditional independence of *intra*-segment residues. It is explicitly chosen to allow the modeling of correlations among positions within a segment of the type described in Chapter 2. Thus the terms for individual segments may take on arbitrary form, and may depend on global properties of the

[1] I will restrict attention to the 3-state problem where $\forall i \; T_i \in \{H, E, L\}$, although generalizations may be desirable and are easily accommodated.

segment as a whole (such as hydrophobic moment or helix dipole) beyond properties of individual residues.

4.2.3 Prior

Given (4.1), it remains to provide a prior distribution $P(\mathcal{S}) = P(m, S, T)$ to completely specify the joint distribution $P(R, \mathcal{S})$. The approach taken here is to factor $P(S, T \mid m)$ as a semi-Markov process:

$$P(m, S, T) = P(m) \prod_{j=1}^{m} P(T_j \mid T_{j-1}) P(S_j \mid S_{j-1}, T_j) \tag{4.2}$$

yielding the joint distribution:

$$P(R, m, S, T) = P(m) \prod_{j=1}^{m} P(R_{[S_{j-1}+1:S_j]} \mid m, S, T) P(T_j \mid T_{j-1}) P(S_j \mid S_{j-1}, T_j) \tag{4.3}$$

Under this model, each segment type depends only on its neighbors, and the segment length distributions $P(S_j \mid S_{j-1}, T_j)$ are conditioned on segment type. This allows explicit modeling of the differences in length observed in experimental protein structures (see Figure 4.2). $P(m)$ may be arbitrary here, but affects the computational complexity of inference as described in Chapter 7. This factorization of $P(S, T)$ produces a model closely related to the class of *hidden semi-Markov* or *semi-Markov source* models discussed in (Russell and Moore, 1985; Levinson, 1986; Rabiner, 1989) for applications in speech recognition. In the speech recognition literature however, observations during a given state occupancy are typically modeled as independent and identically distributed (*iid*). As demonstrated in Section 4.3 below, the ability to model both non-independence and non-identity of distributions is a major motivation for this segment-based approach. A more detailed comparison of the relation between HMMs, HSMMs, and segmentation models is given in Chapter 6.

The factorization given by (4.2) is only one of many possible choices for the prior

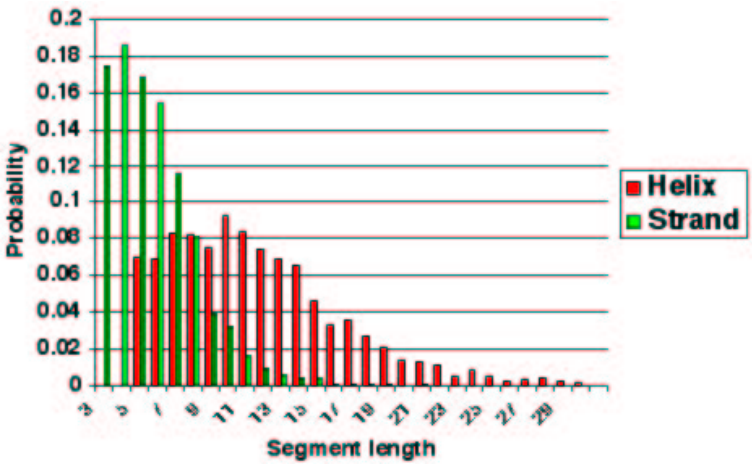


Figure 4.2: Empirical length distribution of observed structural segments for α-helices (red) and β-strands (green). Distributions are calculated from the structural database described in Chapter 9.

$P(\mathcal{S})$. Alternate priors are discussed in some detail in Chapter 6.

4.3 Example models for α-helices and β-strands

I now introduce a set of concrete probability models for secondary structure segments which will serve to demonstrate the approach developed in Section 4.2. The goal is to choose a specific form of the segment likelihood $P(R_{[S_{j-1}+1:S_j]} \mid S, T)$ which captures the core aspects of protein secondary structure formation described in Chapter 2: amino acid propensities, helical capping signals, hydrophobicity patterns, and side chain interactions. In other words, the goal is to develop probabilistic *models* for protein structural segments. For example, the function $P(R_{[i:j]} \mid i, j, H)$ provides the likelihood of the subsequence $R_{[i:j]}$ under the assumption that a helix begins at position i and ends at position j. Given such a segment likelihood for each structural class $\{H, E, L\}$, computing the likelihood of a sequence under any given structural assignment is trivially done by evaluating the joint distribution (4.1). Here I provide the specific form for a set of such segment likelihoods. These segment models will be used in Chapter 9 to evaluate the methodology developed here. Development of alternative segment models is discussed briefly in Chapter 10.

4.3.1 Helix model

As discussed in Chapter 2, the presence of correlated side chain mutations in α-helices has been well studied, deriving from both environmental constraints and stabilizing side chain interactions. Because the α-helical and β-strand backbone conformations yield different periodicities (Figure 2.7), these correlations can provide important clues for identifying the secondary structure from sequence. Another important source of information for identifying α-helical segments are the helix capping signals discussed in Section 2.2.1. This capping effect results in amino acid distributions at end-segment positions which differ significantly from those in other positions of the helix (Table 2.1).

The goal is to develop a helical segment model which captures such position-specific preferences and probabilistic dependence of intra-segment residues, in addition to standard amino acid propensities. The model chosen must also account for helices of various lengths. The following form can be used to capture all of these sources of information:

$$P(R_{[S_{j-1}+1:S_j]} \mid S_{j-1}, S_j, H) = \prod_{i=S_{j-1}+1}^{S_{j-1}+\ell_N^H} P^H_{N_{i-S_{j-1}}}\left(R_i \mid R_{[S_{j-1}+1:i-1]}\right) \times \prod_{i=S_{j-1}+\ell_N^H+1}^{S_j-\ell_C^H} P^H_I\left(R_i \mid R_{[S_{j-1}+1:i-1]}\right) \times \prod_{i=S_j-\ell_C^H+1}^{S_j} P^H_{C_{S_j-i+1}}\left(R_i \mid R_{[S_{j-1}+1:i-1]}\right) \tag{4.4}$$

Here ℓ_N^H indicates the length of the helix N_{cap} model, N_i and C_i indicate the i^{th} position from the N- and C-termini respectively; and I indicates an internal (non-cap) position. Figure 4.3 shows graphically how this model is applied to the particular amino acid subsequence of the helix in Figure 2.6: the first product in (4.4) models the distribution of amino acids at each of the first N-terminal positions ($N_{cap}, N_1, N_2, N_3, \ldots$), and similarly the last term for the C-terminal positions ($\ldots, C_3, C_2, C_1, C_{cap}$), while the middle term models all internal positions as identically distributed but dependent.

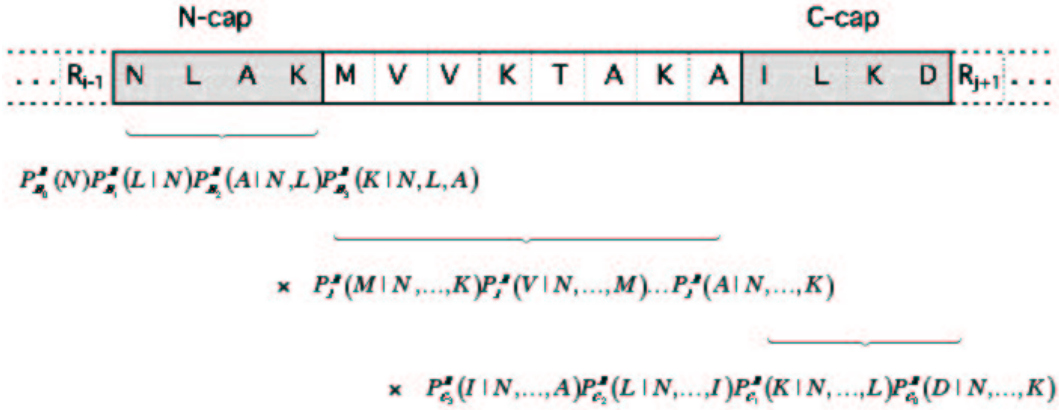


Figure 4.3: Evaluation of the α-helix segment model for a particular amino acid subsequence. Grayed areas are the N- and C-capping positions, specified by distinct amino acid probability distributions. Internal positions are modeled as identically distributed but dependent. Throughout, amino acid distributions are conditioned on neighboring residues according to known helical side chain interactions, as described in Section 4.3.1.

Choosing the length of the helix cap models amounts to deciding which positions of the helix have position-specific propensities which differ significantly from internal positions. Based on the observations discussed in Chapter 2 with regard to capping effects, hydrogen bonding patterns, and dipole interactions, the first and last 4 positions are the most likely candidates to consider. Figure 2.5 shows distributions for these positions observed in the database compiled in Chapter 9. Table 2.1 shows the statistical deviance between the amino acid distribution at each end-segment position and the amino acid distribution at internal positions. The strongest signal appears in the first two positions of the helical N-terminus (N_1 and N_2). The positions chosen for the experiments described in Chapter 9 are highlighted (so that $l_N^H = 4$, $l_C^H = 1$). It is worth noting that such information is inherently difficult to include in the window-based prediction methods described in Chapter 3, which must scan a residue across each position in the window in turn.

Equation (4.4) does not provide the exact intra-segment residue dependencies $P_i^H\left(R_i \mid R_{[j:i-1]}\right)$ in the model. This is again an issue of model choice. Because modeling the full joint distribution of 3 or more amino acids leads to an explosion in

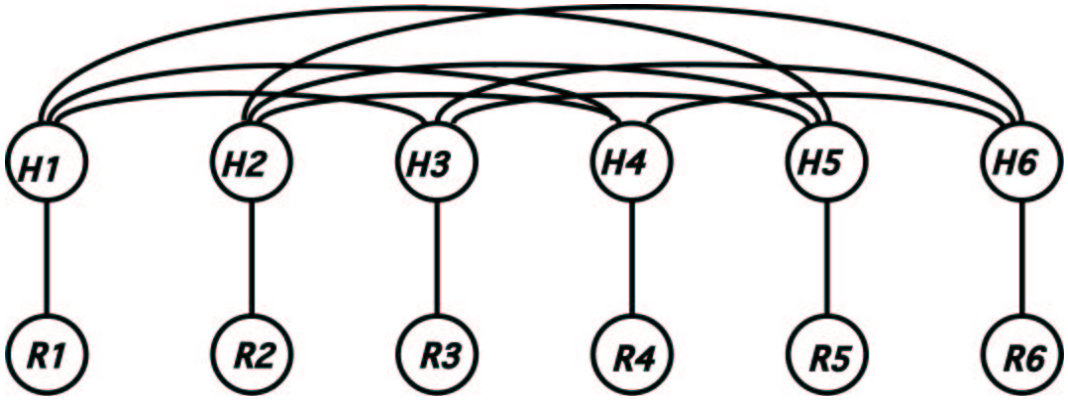


Figure 4.4: Graphical model (Whittaker, 1990) representing the conditional independence structure for amino acids in an example α-helix. R_i are the amino acids, and H_i are their associated hydrophobicity classes as assigned by (Klingler and Brutlag, 1994). The model provides for dependence among the hydrophobicity classes at appropriate periodicity, allowing the amino acid distributions to be modeled as *conditionally* independent, thus reducing the dimensionality of the model.

the number of parameters which quickly outstrips the amount of experimental data available for estimation, it is desirable to model the intra-segment dependencies in some more restrictive fashion. The following model, used in the experiments of Chapter 9 and elsewhere (Schmidler et al., 2000), attempts to capture the hydrophobicity patterns displayed in Figures 2.6 and 2.7:

$$P_i^H\left(R_i \mid R_{[j:i-1]}\right) = P_i^H\left(R_i \mid h_i\right) P_i^H\left(h_i \mid h_{i-2}, h_{i-3}, h_{i-4}\right) \tag{4.5}$$

where $h_i \in \{\text{hydrophobic, neutral, hydrophilic}\}$ indicates the hydrophobicity class of residue R_i as assigned by (Klingler and Brutlag, 1994). In other words, dependency between positions is modeled using a reduced alphabet in order to avoid combinatorial explosion of parameters. Figure 4.4 provides a graphical model (Whittaker, 1990) representation of the dependency structure given by (4.5). This form of the distribution captures explicitly the previously described intra-segment residue correlations corresponding to the periodicity of an α-helix by conditioning the probability of a particular residue on the $i-4^{th}$, $i-3^{rd}$, and $i-2^{nd}$ residues. Internal positions are therefore modeled as identically distributed, but dependent. This of course is not the

only possible choice for this residue dependency, and approaches to optimizing this model are suggested in Chapter 10.

4.3.2 β-strand and loop/coil models

The general form given by (4.4) is convenient for modeling variable-length segments, and can be retained when developing β-strand and loop segments as well. However the utility of distinguishing end-capping residues in β-strands and loops is less obvious than in the case of α-helices. For example, there are no positions in β-strands which are physically analogous to the capping positions of α-helices. Table 2.1 shows the statistical deviance of segment-end positions of both β-strands and loop/coil segments, and there is little evidence of variation in these positions. Accordingly, the models used in Chapter 9 set $\ell_N^E = 1$, $\ell_C^E = 1$, $\ell_N^L = 2$, $\ell_C^L = 1$. Figure 2.5 shows the distributions at these positions.

Another difference between the models for α-helices, β-strands, and loops lies in the form of the intra-segment residue dependency. The dependency given by (4.5) reflects the intra-segment correlations induced by the underlying backbone-side chain geometry of α-helices. These correlations are expected to differ for β-strands (Figure 2.7) and loops/coils. To account for these differences, conditioning is done on residues $i-1$ and $i-2$ for both β-strands and loops. Again, the issue of selecting the best form for such interactions is discussed briefly in Chapter 10.

4.4 Secondary structure prediction

Assuming a probability model given by (4.1) in conjunction with a set of segment models such as (4.4,4.5), we require a method for deriving structure predictions for a new protein sequence R. This is accomplished by inferring the secondary structure assignment parameters (m, S, T) for R. In the Bayesian framework developed here, such inferences are based on the posterior distribution over parameters $P(m, S, T \mid R)$. The goal is thus to find (m, S, T) such that $P(m, S, T \mid R)$ is maximized. There are

two plausible definitions for predictors which maximize $P(m, S, T \mid R)$:

$$Struct_{MAP} = \arg \max_{(m,S,T)} P(m, S, T \mid R) \quad (4.6)$$

$$Struct_{Mode} = \{\arg \max_{T} P(T_{R_{[i]}} \mid R)\}_{i=1}^{n} \quad (4.7)$$

where $P(T_{R_{[i]}} \mid R)$ denotes the marginal posterior distribution over structural types at a single position i in the sequence:

$$P(T_{R_{[i]}} \mid R) = \sum_{(m,S,T)} P(m, S, T \mid R) \mathbf{1}_{\{T_{R_i}=t\}} \quad (4.8)$$

and $Struct_X$ is a segmentation of R.

$Struct_{MAP}$ defined by (4.6) provides the *maximum a posteriori* (MAP) segmentation of a sequence, the segmentation which maximizes the joint posterior probability $P(m, S, T \mid R)$. However, a large number of sub-optimal segmentations may exist which also contribute significant probability mass to the posterior. In addition, the most common accuracy measure for protein secondary structure prediction is the Q_3 value, the percentage correct on a *per-residue* basis (see Chapter 9). Thus the MAP segmentation is not as desirable as $Struct_{Mode}$ (4.7), the predictor defined by the sequence of marginal posterior modes, those structural assignments which maximize $P(T_{R_{[i]}} \mid R)$ at each position i. Note that calculation of $P(T_{R_{[i]}} \mid R)$ in (4.7) involves marginalization over all possible segmentations, while (4.6) involves maximization over this space. Clearly calculation by direct enumeration is infeasible, and more efficient algorithms are required. Algorithms for computation of (4.6) and (4.7) are given in Chapter 7.

It is worth noting that the Bayesian framework can easily incorporate prior knowledge about regions or positions in the sequence if such is available. The algorithms provided in Chapter 7 may be easily modified to calculate probabilities *conditional* on certain positions or segments taking on known conformations. This might be the case if experimental evidence exists such as circular dichroism data or foot-printing experiments, or if highly significant motif hits occur on the sequence and provide structural information, for example with helix-turn-helix DNA binding motifs. Again,

such information is inherently difficult to include in most existing secondary structure prediction methods.

4.5 Concluding remarks

In this Chapter I have developed a Bayesian framework for protein structure prediction based on a segment-decomposition of the joint probability distribution. I have defined a set of probabilistic models for protein structural segments, and shown how prediction of protein secondary structure may be achieved under this general class of models. In Chapter 9 I will subject the methods developed here to experimental validation, and show that they perform at the level of the best existing algorithms for secondary structure prediction. Before proceeding to this evaluation however, I will introduce a generalization of this framework in Chapter 5, which allows incorporation of non-local interactions. Chapter 5 shows how the framework introduced here may be applied beyond the realm of secondary structure prediction to obtain predictions of tertiary contacts and β-sheet topology.

Chapter 5

Bayesian Modeling of Non-Local Interactions

In Chapter 4, I introduced the basic modeling framework developed in this dissertation. This framework formulates protein structure prediction in terms of Bayesian inference and provides a class of joint probability models factored by structural segments. I then showed how this framework can be applied in practice by providing a concrete set of models for secondary structure segments. Algorithms for inference of secondary structure from sequence will be discussed in Chapter 7.

In this Chapter, I demonstrate the generality of this framework by extending the class of segment-based models to include non-local interactions in protein sequences. This enables treatment of a significantly broader class of protein structure prediction problems than standard secondary structure prediction. I show that probability models for non-local interactions can be incorporated into the Bayesian framework in a conceptually simple manner by the introduction of joint-segment models. I demonstrate this approach by developing models for correlated mutations in neighboring β-strands of a β-sheet. I show how the Bayesian framework naturally incorporates these interactions into secondary structure prediction, and also yields predictors for tertiary β-strand contacts and β-sheet topology.

The methods developed here provide a rigorous framework for synthesizing these

diverse types of information in an optimal way, and hence obtaining structure predictions which combine all available information. These benefits do not come without a price however, as computation in this more general class of models is significantly more difficult. Methods for computing with the joint-segment models introduced here will be developed in Chapter 8.

5.1 Motivation

A fundamental assumption of the class of models described by (4.1) is the conditional independence of amino acids which occur in distinct segments. This assumption enables the exact calculation of posterior probabilities using the recursions provided by (7.1-7.7), and hence the predictors $Struct_{MAP}$ and $Struct_{Mode}$ given by (4.6) and (4.7) can be computed efficiently.

However, this assumption is clearly violated in the case of protein sequences. Non-local sequence dependencies are introduced by evolutionary pressure to maintain non-local interactions important for for protein folding of the type described in Chapter 2. I define an interaction to be *non-local* if it involves amino acids occurring in distinct structural segments. By expanding the set of segment classes as needed, all non-local interactions in protein folding may therefore be represented.

For example, a β-sheet consists of multiple β-strands linked by backbone hydrogen bonds (Figure 2.3). Correlation between amino acids on different strands within a β-sheet was described in Section 2.2.2. Hence β-sheets form a major structural motif in proteins which relies on interaction between sequentially distant segments, or *non-local* interaction, to form a stable native fold. Other common examples include disulfide bonds and coiled coils (Section 2.2.2), and the presence of correlated mutations in such motifs is well known (Krylov et al., 1994; Lifson and Sander, 1980; Wouters and Curmi, 1995; Hutchinson et al., 1998). Again, such dependencies arise through evolutionary pressure as a consequence of common chemical environment or stabilizing side chain interactions.

It has been often been suggested that the difficulty of capturing such non-local patterns in protein sequences may be responsible for the low accuracy typically achieved

by secondary structure prediction algorithms in identifying β-strands. In the next section, I show how the framework developed in Chapter 4 may be extended to account for such inter-segment residue correlations by introducing joint segment probability models.

5.2 Segment interactions and β-sheet topologies

The important new idea introduced in this Chapter is that the class of stochastic segment models described by (4.1) can be generalized to a much larger class of models involving *segment interactions*. We may write a joint distribution over the set of *interacting segmentations* in the following form:

$$P(R, \mathcal{S}, \mathcal{I}) \propto P(\mathcal{S}, \mathcal{I}) \prod_{i=1}^{p} P(\{R_{[s_j:e_j]}\}_{S_j \in \mathcal{H}_i} \mid \mathcal{S}, \mathcal{I}) \tag{5.1}$$

which factors by conditional independence of *sets of interacting* segments. Here modeling of inter-segment sequence dependencies is achieved by introducing *joint-segment likelihoods*, replacing the terms

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

for two interacting segments S_j and S_k in the product of (4.1) above with a joint term:

$$P(R_{[s_j:e_j]}, R_{[s_k:e_k]} \mid S_j, S_k) \tag{5.2}$$

Hence positions R_i in different segments may be made conditionally *dependent* by introducing an interaction between the two segments, and 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.

In order to define a joint distribution of the form (5.1), we must first define what is meant by a *segment interaction*. In this Chapter I will do so by developing an example, developing joint-segment models for β-sheets. I begin by introducing models

for a restricted class of β-sheets known as β-hairpins, and then extend this approach to arbitrary β-sheet topologies. Chapter 6 provides a more formal treatment of the general class of probability models on segment interactions represented by (5.1) and introduced in this dissertation.

5.2.1 β-hairpins

Representing β-hairpins

In order to demonstrate the approach, we first consider a simple example of inter-segment interactions: the pairing of two β-strands in an anti-parallel orientation to form a β-hairpin.

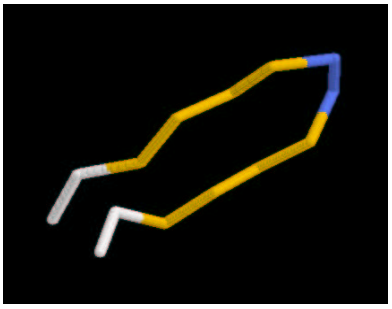
A β-hairpin is an anti-parallel β-sheet made up of two strands separated by a short hairpin loop (see Figure 5.2.1a) and linked by inter-strand hydrogen bonds. β-hairpins have been well-studied, and detailed classifications exist (Sibanda et al., 1989). In order to capture the interaction between strands which makes up a β-hairpin, we must extend the segmentation notation $\mathcal{S} = (m, S, T)$ to include additional parameters.

For simplicity, we may consider a β-hairpin to be any anti-parallel β-sheet involving two β-strands separated by a single segment. A simple representation of such a β-hairpin involves an interaction between only two segments[1], the participating β-strands S_i and S_j with $j = i + 2$. In order to specify the interaction between S_i and S_j, we must specify which positions of each segment interact[2]. In the case of β-hairpins, this is particularly simple. The relative orientation of the two strands is known (anti-parallel), and if we assume that interaction among positions in a β-hairpin occurs in one contiguous stretch (excluding for the moment non-contiguous interactions such as β-bulges)[3], we need only specify the register of the C-terminal strand (S_j) relative to the N-terminal strand (S_i). We specify this interaction by introducing the parameters $h_{i,j}, h_{j,i}, \ell_{i,j}$, with the following interpretations:

[1] A more complicated model might include the hairpin loop segment, in order to model correlations between strand and loop residues.

[2] *Interaction* between positions here does not necessarily mean physical interaction, but more generally a (conditional) dependence in the structure of the joint probability distribution.

[3] More generally, interacting positions may be specified by a contact matrix of binary interaction indicators, but we do not pursue this here.



(a)

Figure 5.1: (a) β-hairpin from bovine pancreatic phospholipase (1bp2). (b) Parameterization of β-hairpin segment interaction.

(i) $h_{i,j}$ is the first (N-terminal) position of segment i which interacts with a position on segment j

(ii) $\ell_{i,j}$ is the number of (contiguous) interacting positions in the interaction of segment i with segment j

A simple example of this notation is given in Figure 5.2.1b. The notation h is used to suggest hydrogen bonding, one determinant of which positions interact in β-sheets. A β-hairpin interaction is thus specified by:

$$I = (\mathcal{H}, h) = (\{S_i, S_{i+2}\}, \{h_{i,i+2}, h_{i+2,i}, \ell_{i,i+2}\}) \tag{5.3}$$

and a set of p β-hairpins for a sequence denoted by:

$$\mathcal{I} = \left\{I^i\right\}_{i=1}^{p} = \left\{(\mathcal{H}^i, h^i)\right\}_{i=1}^{p} \tag{5.4}$$

We are now in a position to define probability models for segmentations containing β-hairpin segment interactions, using the parameterization developed in this section.

Joint-segment likelihoods for β-hairpins

Having parameterized a β-hairpin, we now define a joint-segment likelihood for the interacting β-strand segments. We begin with a simple model which accounts only

for pairwise correlation between cross-strand neighboring residues:

$$P(R_{[s_i:e_i]}, R_{[s_j:e_j]} \mid \mathcal{S}, \mathcal{I}) = \prod_{\substack{k = s_i,\ldots,h_{i,j}-1 \\ s_j,\ldots,h_{j,i}-1 \\ h_{i,j}+\ell_{i,j},\ldots,e_i \\ h_{j,i}+\ell_{i,j},\ldots,e_j}} P(R_k \mid \mathcal{S}, \mathcal{I}) \prod_{k=0}^{\ell_{i,j}-1} P(R_{[h_{i,j}+k]}, R_{[h_{j,i}+\ell_{i,j}-1+k]} \mid \mathcal{S}, \mathcal{I}) \quad (5.5)$$

Terms in the right-hand product model the joint probability of cross-strand neighboring residues, while terms in the left-hand product account for the N- and C- terminal ends.

The β-hairpin model given by (5.5) is a concrete example of joint-segment likelihoods of the form (5.2). It is also perhaps the simplest possible model which accounts for cross-strand dependency. Notice that each cross-strand residue pair is modeled as *iid*, and no int*ra*-segment dependency is included for any position. The intra-segment dependency and position-dependent probability models developed for β-strands in Chapter 4 may of course be added here.

Priors for β-hairpin interactions

In order to complete the joint distribution (5.1) we must also specify the prior distribution $P(\mathcal{S}, \mathcal{I})$. It may now be desirable to specify prior distributions on segment locations, lengths, and types within a segment interaction *jointly*:

$$P(S_j, S_k \mid S_{j-1}, S_{k-1}, T_j, T_k) P(T_j, T_k) \quad (5.6)$$

It is also necessary to specify a prior distribution over the topology parameters themselves:

$$P(\mathcal{I} \mid \mathcal{S}) = P((\{S_i, S_{i+2}\}, \{h_{i,i+2}, h_{i+2,i}, \ell_{i,i+2}\})_{i=1}^{p} \mid \mathcal{S}) \quad (5.7)$$

We will return to the issue of priors in the next section.

Stochastic Segment Interaction Model for β-hairpins

Restricting ourselves to only β-hairpin interactions then, the joint distribution over segmentations may be written as:

$$P(R, \mathcal{S}, \mathcal{I}) = P(\mathcal{S}, \mathcal{I}) \prod_{i=1}^{m-2p} P(R_{[s_i:e_i]} \mid \mathcal{S}, \mathcal{I}) \prod_{i=1}^{p} P(R_{[s_{H_i}:e_{H_i}]}, R_{[s_{H_i+2}:e_{H_i+2}]} \mid \mathcal{S}, \mathcal{I}) \quad (5.8)$$

where as before $\mathcal{S}$ is a set of segments, and $\mathcal{I}$ is a set of hairpin interactions of the form (5.3). Recall also the restriction that the sets of interacting segments are disjoint, so that by definition no strand may participate in > 1 β-hairpin simultaneously.

Experiments using the model given by (5.8) to predict β-hairpins in real protein sequences are described in Chapter 9.

5.2.2 β-sheets

Representing arbitrary β-sheets requires further extensions to the segmentation parameters, beyond those introduced for the special case of β-hairpins.

We adopt the following specification of segment interactions for representing strand-pairing and β-sheet formation:

(i) k is the number of segments (β-strands) participating in the sheet

(ii) $h_{i,j}$ are the parameters specifying the interaction between segments i and j. For β-sheets, $h_{i,j}$ will be non-empty for only the $2k-2$ neighbors, where we require that each segment in the sheet interact with ≥ 1 and ≤ 2 other segments (referred to as Left and Right partners).

(iii) A β-sheet I is therefore specified as:

- A set of segment indices $\mathcal{H} = \{H_j\}_{j=1}^{k}$
- For each segment H_j in $\mathcal{H}$, the parameters
  - i) Left interaction $= (n_{H_j,l}, a_{H_j,l}, b^N_{H_j,l}, b^C_{H_j,l}) = h_{H_j,n_l}$
  - ii) Right interaction $= (n_{H_j,r}, a_{H_j,r}, b^N_{H_j,r}, b^C_{H_j,r}) = h_{H_j,n_r}$

Symbol	Values	Description
$n_{i,l}$	$\{1,\ldots,m\}$	Segment number of Left interaction partner of segment i.
$a_{i,l}$	$\{1,-1\}$	Parallel or anti-parallel orientation of segment i with Left interaction partner.
$b_{i,l,N}$	$\{S_{i-1}+1,\ldots,S_i\}$	First (N-terminal) position in segment i which interacts with Left interaction partner positions.
$b_{i,l,C}$	$\{b_{i,l,N},\ldots,S_i\}$	Last (C-terminal) position in segment i which interacts with Left interaction partner positions.

Table 5.1: Parameters used to specify topology of a β-sheet.

Figure 5.2: Put a figure here showing (a) a β-hairpin and (b) a 3-stranded parallel β-sheet along with the associated parameters for specifying each.

(iv) Interaction parameters $h_{i,j} = (j, a_{i,j}, b^N_{i,j}, b^C_{i,j})$ represent the interaction of segment i with segment j. Here

- $a_{i,j}$: Specifies the relative orientation of i with respect to j ($a_{i,j} = 1$: parallel; $a_{i,j} = -1$: anti-parallel)
- $b^{N/C}_{i,j}$: Specifies the N/C-terminal position of segment i which interacts with positions in segment j. Interacting positions are assumed to be contiguous (e.g. no β-bulges).
- For β-sheets the following symmetries are imposed:
  - $a_{i,j} = a_{j,i}$
  - $\ell_{i,j} = (b^C_{i,j} - b^N_{i,j}) = (b^C_{j,i} - b^N_{j,i}) = \ell_{j,i}$, the interacting subsequences are of equal length.

This parameterization is slightly redundant, but will prove convenient for notational purposes. Parameters are summarized in Table 5.1. Examples of this parameterization are given in Figure 5.2. Note that edge strands of a sheet have only one partner. Further restrictions on allowable sheet topologies (such as requiring all neighbors to have the same orientation a), may be introduced via the probability model, as described in Section 5.2.3.

Joint segment models for β-sheets

The model described in Section 5.2.1 can be extended to arbitrary β-sheet topologies in a relatively straightforward manner. We proceed by specifying the β-sheet model in terms of pairwise interactions between neighboring strands. More sophisticated models specifying the joint distribution of residues on > 2 strands may be desirable, but this approach suffices for the current demonstration. The joint distribution for sheet interaction I^i is then given by:

$$P(\{R_{[s_{H_j}:e_{H_j}]}\}_{H_j \in \mathcal{H}_i} \mid \mathcal{S}, \mathcal{I}) = \left[\prod_{j=2}^{k_i - 1} \prod_{i=s_{H_j}}^{e_{H_j}} P(R_{[i]}) \right]^{-1} \times$$

$$\prod_{j=1}^{k_i - 1} \left[\begin{array}{c} \prod_{\substack{s_{H_j}, \ldots, b^N_{H_j, H_{j+1}} - 1 \\ i = \substack{s_{H_{j+1}}, \ldots, b^N_{H_{j+1}, H_j} - 1 \\ b^C_{H_j, H_{j+1}} + 1, \ldots, e_{H_j} \\ b^C_{H_{j+1}, H_j} + 1, \ldots, e_{H_{j+1}}}}} P(R_{[i]} \mid \mathcal{S}, \mathcal{I}) \prod_{i=0}^{\ell_{H_j, H_{j+1}} - 1} P(R_{[b^N_{H_j, H_{j+1}} + i]}, R_{[ptnr]} \mid \mathcal{S}, \mathcal{I}) \end{array} \right] \quad (5.9)$$

where

$$ptnr = \begin{cases} b^C_{H_{j+1}, H_j} - i & \text{if} \quad a_{i,j} = -1 \quad \text{(pairing is anti-parallel)} \\ b^N_{H_{j+1}, H_j} + i & \text{if} \quad a_{i,j} = 1 \quad \text{(pairing is parallel)} \end{cases} \quad (5.10)$$

As with the β-hairpin model, this model captures only the simplest possible dependency structure for amino acids within a β-sheet.

5.2.3 Sheet interaction priors

Having specified the joint-segment likelihood for β-hairpins and general β-sheets, we return to the issue of priors over segment interaction parameters.

The same priors on segmentations discussed in Chapter 4 may be applied here.

In this case, a uniform prior:

$$P(\mathcal{S}, \mathcal{I}) \propto 1 \tag{5.11}$$

will yield the maximum likelihood segmentation with interactions. Another approach is to specify the prior on interactions conditionally:

$$P(\mathcal{S}, \mathcal{I}) = P(\mathcal{I} \mid \mathcal{S}) P(\mathcal{S}) \tag{5.12}$$

using e.g. one of the priors $P(\mathcal{S})$ described in Chapter 4. To make this more concrete, consider the interaction parameterization developed in previous sections for modeling β-sheets. We might consider all interactions for a given segmentation to be equally likely:

$$P(\mathcal{I} \mid \mathcal{S}) = 1/c(\mathcal{S}) \tag{5.13}$$

where $c(\mathcal{S})$ is the number of possible sets of segment interactions $\mathcal{I}$ for a given set of segments. Using the parameterization of the previous sections, we may calculate this as follows:

Let $\mathcal{S}$ be a segmentation with k β-strand segments $\{S_j\}_{j=1}^k$ with lengths $\{\ell_j\}_{j=1}^k$. Then the number of possible sets of β-sheets is given by:

$$c(\mathcal{S}) = \sum_{p \in \mathcal{P}(k)} \prod_{U \in p} \sum_{\sigma \in S_{|U|}} (2^{|U|-1} - 1) \prod_{i=1}^{|U|-1} (\ell_{U_{\sigma^{-1}(i)}} + \ell_{U_{\sigma^{-1}(i+1)}} - 1) \tag{5.14}$$

where:

- $\mathcal{P}$ is the set of partitions of the integers $1, \ldots, k$, corresponding to the various groupings of strands into interaction sets (β-sheets)

- U are subsets in partition p, each corresponding to a β-sheet

- S_n is the group of permutations of n elements, each corresponding to a possible topology connecting β-strands within the β-sheet

- $(2^{|U|-1} - 1)$ is the number of possible orientations of strands in a given topology

- $(\ell_1 + \ell_2 - 1)$ is the number of possible registers for aligning two neighboring β-strands, requiring ≥ 1 pair of interacting positions

Clearly (5.14) grows very rapidly in k. This indicates that prior (5.13) has a substantially different effect than (5.11). In particular, the marginal distribution $P(\mathcal{S})$ of (5.11) is highly biased towards segmentations containing many β-strands. In contrast, (5.13) preserves the marginal distribution $P(\mathcal{S})$ by requiring

$$\sum_{\mathcal{I}} P(\mathcal{I} \mid \mathcal{S}) = 1 \;\; \Rightarrow \;\; \sum_{\mathcal{I}} P(\mathcal{S}, \mathcal{I}) = P(\mathcal{S}) \tag{5.15}$$

In other words, (5.11) produces a sensible noninformative prior over interactions, but marginally produces a highly biased prior for the segments $\mathcal{S}$ themselves, whereas (5.13) specifies a prior which is highly biased against any interactions $\mathcal{I}$, but marginally sensible on segmentations.

In practice, (5.14) is infeasible to compute for large sequences. However, given this structure of the interaction space, we may define another sensible prior which also preserves (5.15):

$$P(\mathcal{I} \mid \mathcal{S}) = P(p \mid \mathcal{S}) \prod_{U \in p} P(\sigma \mid U) P(a_U) \prod_{i=1}^{|U|-1} P(align(U_{\sigma(i)}, U_{\sigma(i+1)}))$$

Here

$$P(a_U) = P(\{a_{\sigma^{-1}(i),\sigma^{-1}(i+1)}\}_{i=1}^{|U|-1})$$

is the joint distribution of all β-strand orientations in a β-sheet. Currently we restrict strand pairing in β-sheets to be either all parallel or all anti-parallel, corresponding to the two major classes of β-sheet observed in experimental protein structures:

$$P(a_U) = \begin{cases} q & a_{i,j} = 1 \quad \forall a_{i,j} \in \{a_{\sigma^{-1}(i),\sigma^{-1}(i+1)}\}_{i=1}^{|U|-1} \\ 1-q & a_{i,j} = -1 \quad \forall a_{i,j} \in \{a_{\sigma^{-1}(i),\sigma^{-1}(i+1)}\}_{i=1}^{|U|-1} \\ 0 & otherwise \end{cases}$$

With q representing the relative frequency of parallel vs. anti-parallel β-sheets. q

may be estimated from the databases described in Chapter 9. The remaining terms are specified as follows:

$$\begin{aligned} P(p \mid \mathcal{S}) &= 1 && (5.19)\\ P(\sigma \mid U) &= |U|!^{-1} && (5.20)\\ P(align(i,j)) &= P(b^N_{i,j})P(b^C_{i,j})P(b^N_{j,i})P(b^C_{j,i}) \end{aligned}$$

where we set:

$$P(b^N_{i,j}) = \begin{cases} \frac{1}{3} & (b^N_{i,j} - S_i) \leq 3 \\ 0 & otherwise \end{cases}$$

This states that all topologies of strands are equally likely, and for two paired strands the first (last) interaction or hydrogen bond is uniform over the first (last) 3 positions of the strand. The latter assumptions, including uniformity and independence of the offsets, is clearly untrue and it may be desirable to estimate these distributions from data.

Hence our prior on a set of sheet interactions is of the form:

$$\begin{aligned} P(\mathcal{I} \mid \mathcal{S}) &= \prod_{j=1}^{p} P(I_j \mid \mathcal{S}) && (5.22)\\ &= p(a_{I_j}) \prod_{j=1}^{p} (k_j!)^{-1} \prod_{i=1}^{k_j - 1} P(b^N_{H_{j,i},H_{j,i+1}})P(b^C_{H_{j,i},H_{j,i+1}})P(b^N_{H_{j,i+1},H_{j,i}})P(b^C_{H_{j,i+1},H_{j,i}}) \end{aligned}$$

recall that $p = |\mathcal{I}|$ so $\{I_i\}_{i=1}^{p}$ is the set of β-sheets.

For a single β-hairpin as shown in Figure 5.2a, this reduces to:

$$\begin{aligned} P(\mathcal{I} \mid \mathcal{S}) &= P(\mathcal{H} = (j,k), b^N_{j,k}, b^C_{j,k}, b^N_{k,j}, b^C_{k,j}, a_{j,k} = a_{k,j} = -1 \mid \mathcal{S}) \\ &= p(a_{I_1} = -1)\frac{1}{2}P(b^N_{i,j})P(b^C_{i,j})P(b^N_{j,i})P(b^C_{j,i}) && (5.23) \end{aligned}$$

Non-conditional interaction priors: In some cases it may be sensible to specify the joint distribution of $(\mathcal{S}, \mathcal{I})$ directly, allowing the prior on $\mathcal{S}$ to reflect the nature of interactions specified in $\mathcal{I}$. For example, using a semi-Markov prior on segmentations

(4.2) we may wish to model the length distributions of interacting segments jointly, as in (5.6). For the β-hairpin example (Figure 5.2a), this produces:

$$\begin{aligned} P(\mathcal{I},\mathcal{S}) &= \frac{P(\ell_j,\ell_k)}{P(\ell_i)P(\ell_j)} \prod_{i=1}^{m} P(T_i \mid T_{i-1}) P(\ell_i) \times \\ & \quad p(a_{I_1} = -1) \prod_{j=1}^{p} \frac{1}{2} P(b^N_{i,j}) P(b^C_{i,j}) P(b^N_{j,i}) P(b^C_{j,i}) \end{aligned} \tag{5.24}$$

5.2.4 More sophisticated β-sheet models

Distinguishing residues pairs

The model given by (5.9) assumes all residue pairs are *iid*. However, as discussed in Chapter 2, there is a well studied distinction between amino acid pairs linked by two surrounding backbone hydrogen bonds, and those pairs whose hydrogen bonds form with the alternate neighboring strand (Wouters and Curmi, 1995; Hutchinson et al., 1998). These differences arise due to different C_β distances for the two different types of pairs. It is straightforward to incorporate this distinction into the model given by (5.9).

Inclusion of other neighbors

Another way in which (5.9) differs from previously developed empirical potentials for β-sheet prediction is through modeling each position as dependent only on immediately adjacent residues. However, the form given by (3.2) may have some advantages, and the dependencies in (5.9) may be extended to include other immediate neighbors as well.

Distinguishing edge strand positions

Another source of information discussed in Chapter 2 is the measurable difference in amino acid propensity which is observed in the edge strands of β-sheets (Minor and Kim, 1994a). This source of information has been lost in previous work using potentials of the form (3.2), because this work does not attempt to predict *β-sheets*, but

rather pairs of β-strand residues. Within the Bayesian framework developed in this dissertation however, the joint distribution is evaluated *conditional* on a particular structural assignment and this distinction may be made between various strands in a β-sheet. This may help introduce further accuracy and strand-pairing specificity into the prediction of β-sheets as described in the next Section.

5.3 Prediction of β-sheet topology and tertiary contact maps

In the previous section I defined joint-segment probability models for β-sheets. Section 5.2 defined joint distributions over the space of segmentations involving sheet interactions via (5.1). Here I show how predictive quantities are derived under this more general class of models, in a fashion analogous to Section 4.4. Computation of these quantities is more difficult, and will be discussed in Chapter 8. A more general treatment of interaction models is given in Chapter 6.

5.3.1 Secondary structure prediction

Prediction of secondary structure under the generalized framework proceeds exactly as in Chapter 4. In a direct parallel to predictors (4.6,4.7), we define

$$Struct_{MAP} = \arg\max_{(\mathcal{S},\mathcal{I})} P(\mathcal{S},\mathcal{I} \mid R) \tag{5.25}$$

$$Struct_{Mode} = \{\arg\max_{T} P(T_{R_{[i]}} \mid R)\}_{i=1}^{n} \tag{5.26}$$

where $P(T_{R_{[i]}} \mid R)$ again denotes the marginal posterior distribution over structural types at a single position i in the sequence, but now marginalized over the significantly enlarged space of segmentations *including segment interactions*:

$$P(T_{R_{[i]}} \mid R) = \sum_{(\mathcal{S},\mathcal{I})} P(\mathcal{S},\mathcal{I} \mid R)\mathbf{1}_{\{T_{R_i}=t\}} \tag{5.27}$$

As discussed in Section 5.2.3 the space of possible segmentations has been considerably expanded via the introduction of segment interaction parameters. Thus the arguments made in Chapter 4 against use of the MAP estimate are even more relevant here. Once again, the marginal posterior mode (5.26) serves as an alternative secondary structure predictor which addresses this problem by marginalizing over all possible segmentations. Notice that the discussion of interaction priors (e.g. (5.11) vs. (5.13)) becomes very relevant here. Since (5.11) yields a marginal prior $P(\mathcal{S})$ which is highly biased towards β-strands, it is expected to perform very poorly in secondary structure prediction via (5.26). On the other hand, prior (5.13)) significantly downweights the effects of interaction for any particular β-strand pairing, and so is unlikely to improve secondary structure prediction accuracy. Instead, we may focus on prediction of the contacts themselves.

5.3.2 Contact map prediction

It should be noted that the *maximum a posteriori* segmentation $Struct_{MAP}$ defined by (5.25) has a different form than that of (4.6). This difference is due to the change in definition: specification of a segmentation now requires specification of the interaction parameters defined in Section 5.2 as well. Hence $Struct_{MAP}$ is no longer a set (m, S, T), but a set $(\mathcal{S}, \mathcal{I})$ where $\mathcal{I}$ as previously defined represents the sets of segment interactions and their associated parameters. In other words, $Struct_{MAP}$ provides the *maximum a posteriori* set of segment interactions, in addition to the segment locations and types (m, S, T). In the context of the β-sheet models developed in Section 5.2, this means the MAP values of sheet parameters p, $\{\mathcal{H}\}_{j=1}^{p}$, and $\{h_{i,j}\}_{i,j=1}^{p}$ are specified, identifying the strands involved in sheets and their topology and relative orientations. Hence the $Struct_{MAP}$ predictor includes β-sheet assignment and topology in addition to secondary structure assignment.

We may summarize a set of β-sheet topologies for a given sequence R by considering the associated *predicted β-strand contact map*, defined by a matrix $C_{n\times n} = \{c_{i,j}\}$

where

$$c_{i,j} = \begin{cases} 1 & \text{If } R_i \text{ and } R_j \text{ are paired in a sheet} \\ 0 & \text{otherwise} \end{cases} \tag{5.28}$$

We will denote the contact map derived from $Struct_{MAP}$ as $Contact_{MAP}$:

$$Contact_{MAP} = \{c_{i,j}\}_{i,j=1}^{n} = \begin{cases} 1 & R_i,\, R_j \text{ paired in a sheet in } Struct_{MAP} \\ 0 & \text{otherwise} \end{cases} \tag{5.29}$$

$Contact_{MAP}$ therefore represents an estimator of C using the single highest probability set of β-sheet interactions, as found in $Struct_{MAP}$. However we have already seen reasons why MAP estimation is undesirable. Unfortunately, the marginalization involved in computing $Struct_{Mode}$ via (5.27) at each position fails to retain the interactions of any particular segmentation, and so unlike $Struct_{MAP}$, the predictor $Struct_{Mode}$ does not predict the topology of β-sheets. However, we may also estimate C in an analogous manner to (5.27), by defining the *marginal predicted β-strand contact map*, given by the matrix:

$$Contact_{Mode} = \{c_{i,j}\}_{i,j=1}^{n}$$

where

$$c_{i,j} = P(i \leftrightarrow j) = \sum_{\mathcal{S},\mathcal{I}} P(\mathcal{S},\mathcal{I} \mid R)\mathbf{1}_{\{i \leftrightarrow j\}} \tag{5.31}$$

with $P(\mathcal{S},\mathcal{I})$ given by (5.1). As with $Struct_{Mode}$, the matrix $Contact_{Mode}$ predicts each potential contact marginalized over *all possible segmentations and segment interactions*. Hence it may be expected to provide more accurate predictions of C than $Contact_{MAP}$ by more formal arguments given in Chapter 6.

Calculation of the quantities $Struct_{MAP}$, $Struct_{Mode}$, $Contact_{MAP}$, and $Contact_{Mode}$ defined here is more difficult than the analogous computation in Chapter 4, and is described in Chapter 8. Experiments with using the matrices $Contact_{MAP}$ and $Contact_{Mode}$ as predictors of true β-strand contacts are reported in Chapter 9.

Chapter 6

Stochastic Segment Models and Stochastic Segment Interaction Models

In this chapter I provide a more formal discussion of the class of models introduced in this dissertation. I relate the non-interacting models introduced in Chapter 4 to better-known models such as hidden Markov models and generalizations. The interacting-segment models developed in Chapter 5 are defined carefully and given the name *stochastic segment interaction models* (SSIMs), and their relation to existing models such as stochastic grammars for RNA structure discussed. The issue of segmentation priors is considered in more detail, and the resulting impact on predictors is discussed. This chapter provides a more concise and general statement of the class of models introduced in this dissertation for analysis of biopolymer sequences. Computational issues are left until Chapters 7 and 8.

6.1 Stochastic segment models

6.1.1 Notation

Let $R = (R_1, \ldots, R_n)$ be an observed sequence of random variables taking values in a finite alphabet $\mathcal{A}_R$, and $T = (T_1, \ldots, T_n)$ an associated sequence of unobserved states from a finite alphabet $\mathcal{A}_T$ (see Section 6.1.1). We are particularly interested in the case where R is the sequence of a biological polymer (protein or nucleic acid). $\mathcal{A}_R$ is then the set of 20 naturally-occurring amino acids or 4 nucleotide bases, and the notation is chosen to be suggestive of applications to protein modeling (R for amino acid *residue*). The alphabet $\mathcal{A}_T$ is application dependent, but may 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).

Segmentations: The unobserved state sequence T may also be represented by a sequence of *segments*, defined as $(type, length)$ pairs, obtained by grouping consecutive T_i's of 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 R.

Although $\mathcal{S}$ completely specifies a unique segmentation, it is convenient to introduce the following additional notation:

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

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

(iii) $e_i = s_i + \ell_i - 1$, the sequence position at which segment i ends

The segment locations $\{s_i\}_{i=1}^m$ (or equivalently $\{e_i\}_{i=1}^m$) are referred to as *change points* in the statistical literature. Note the implicit constraints $s_1 = 1$, $e_m = n$, and $s_i = e_{i-1} + 1$ for $i = 2, \ldots, m$.

Segment interactions: In this dissertation I introduce the additional 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 R and a segmentation $\mathcal{S}$, we may define a *segment interaction* I to be a pair $(\mathcal{H}, H)$ where $\mathcal{H} \subset \mathcal{S}$ is a set of segments and H a set of parameters specifying the precise pattern of interaction among them. More formally, let $\mathcal{S} = \{S_i\}_{i=1}^m$ be the segmentation of R, then $I = (\mathcal{H}, H) = (\{S_{\mathcal{H}(i)}\}_{i=1}^k, \{h_i\}_{i=1}^{2^k})$ where $k = |\mathcal{H}| > 1$ and 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*[1], so $\mathcal{H}_i \cap \mathcal{H}_j = \emptyset$ holds $\forall i \neq j$. If we further define any non-interacting segment in $\mathcal{S}$ to be an interaction of size 1 (so $I_j = (\{S_j\}, \emptyset)$ 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, $\mathcal{I}$ may denote the set of β-sheets in a protein, $\mathcal{H}_i$ the set of strands making up the i^{th} sheet, and H_i parameters specifying the relative orientation and registration of neighboring strands. This example is developed in detail in Chapter 5. Similarly, $I \in \mathcal{I}$ might represent a helical bundle or other super-secondary structure.

6.1.2 Stochastic segment models

In Chapter 4 (see also (Schmidler et al., 2000)), I developed a class of probability models defined on segmentations, of the following form:

$$P(R, \mathcal{S}) \propto P(\mathcal{S}) \prod_{j=1}^{m} P(R_{[s_j:e_j]} \mid \mathcal{S}) \tag{6.1}$$

The key assumption of (6.1) is the conditional independence of positions R_i occurring in different segments, given a segmentation $\mathcal{S}$. (Note that *marginally*, the observed sequence R has a complex dependency structure.) The segment likelihoods $P(R_{[s_i:e_i]} \mid \mathcal{S})$ may be of general form. In Chapter 4 this class of distributions

[1] An interaction is a maximal clique in a triangulated graphical model defined on segments of the sequence.

was shown to be particularly appropriate for modeling aspects of protein sequence-structure dependencies and hence for a Bayesian approach to protein structure prediction. When the segmentation prior $P(\mathcal{S})$ in (6.1) is factored appropriately, these models lend themselves to efficient exact calculation of posterior quantities (see Chapter 7).

A slightly less general form of (6.1) is discussed in (Ostendorf et al., 1996) under the name of *stochastic segment models*, and I adopt this terminology here. A number of related models have been developed in the speech recognition and bioinformatics communities. The relations among these models are discussed in Section 6.2.

6.1.3 Stochastic segment interaction models

The class of stochastic segment models described by (6.1) can be generalized to a much larger class of models involving the segment interactions introduced in Section 6.1.1. We may write a joint distribution over the set of *interacting segmentations* in the following form:

$$P(R, \mathcal{S}, \mathcal{I}) \propto P(\mathcal{S}, \mathcal{I}) \prod_{i=1}^{p} P(\{R_{[s_j:e_j]}\}_{S_j \in \mathcal{H}_i} \mid \mathcal{S}, \mathcal{I}) \tag{6.2}$$

which factors by conditional independence of *sets of interacting* segments. We refer to this new class of models as *stochastic segment interaction models* (SSIMs). Here modeling of inter-segment sequence dependencies is achieved by introducing *joint-segment likelihoods*, replacing the terms

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

for two interacting segments S_j and S_k in the product of (6.1) above with a joint term:

$$P(R_{[s_j:e_j]}, R_{[s_k:e_k]} \mid S_j, S_k) \tag{6.3}$$

Hence positions R_i in different segments may be made conditionally *dependent* by introducing an interaction between the two segments, and 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.

This class of models is sufficiently general to capture the significant non-local dependencies in protein sequences; see Sections 6.2.4 and 5.2 for examples. Letting $p = m$ reduces (6.2) to (6.1), so this class of models strictly generalizes those developed previously here and elsewhere.

As described in Chapter 7, many models of the form (6.1) enjoy nice computational properties. In contrast, models of the form (6.2) usually present significant computational difficulties. While the joint distribution (6.2) is easily evaluated for any fixed segmentation $\mathcal{S}$ of R, calculation of relevant predictive quantities under this model rarely permits efficient exact algorithms. Approximation algorithms based on Monte Carlo simulation are developed in Chapter 8.

6.2 Hidden Markov models and stochastic segment models

The class of models described by (6.1) above has close ties to other stochastic sequence models, and it is helpful to make these explicit.

6.2.1 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 T_i once again be the hidden *state* at position i (rather than the type of

the i^{th} segment; see Section 6.1.1), a HMM may be written in the form[2]:

$$P(R,T) = \prod_{i=1}^{n} P(T_i \mid T_{i-1}) P(R_i \mid T_i) \tag{6.4}$$

Rewriting (6.4) in segment form with T_i now denoting the type of segment i (Section 6.1.1), we obtain:

$$\begin{aligned} P(R,\mathcal{S}) &= \prod_{i=1}^{m} \left[P(T_i \mid T_i - 1) P(R_{s_i} \mid T_i) \prod_{j=s_i+1}^{e_i} P(R_j \mid T_i) P(T_i \mid T_i) \right] \\ &= \left[\prod_{i=1}^{m} P(T_i \mid T_{i-1}) P(T_i \mid T_i)^{\ell_i - 1} \right] \times \left[\prod_{i=1}^{m} \prod_{j=s_i}^{e_i} P(R_j \mid T_i) \right] \end{aligned} \tag{6.5}$$

From (6.5) we observe that HMMs are stochastic segment models with the following additional assumptions:

(i) *Geometric lengths:* Lengths of segments of type T follow a geometric distribution with parameter $p = P(T \mid T)$, so

$$P(\ell = k) \propto p^k \tag{6.6}$$

(ii) *Conditional IID:* Observed sequence positions are *conditionally independent* given the hidden state sequence, including those within the same segment:

$$P(R_i | \, \mathcal{S}, R_{j \neq i}) = P(R_i | \, \mathcal{S}) \tag{6.7}$$

Moreover, within a given segment individual positions are also *identically distributed*:

$$P(R_{[s_i,e_i]} \mid S_i) = \prod_{j=s_i}^{e_i} P(R_j \mid T_i) \tag{6.8}$$

Hence HMMs are a special case of stochastic segment models in which the above additional restrictions are imposed.

[2]To simplify notation, throughout we let $P(T_1 \mid T_0)$ to denote $P(T_1)$.

6.2.2 Hidden semi-Markov models

Assumption (6.6) has been recognized as inappropriate for applications in speech recognition, leading to the development of "explicit state duration density" HMMs (Russell and Moore, 1985; Levinson, 1986; Rabiner, 1989). Such models may be written in the form:

$$P(R, \mathcal{S}) = \left[\prod_{i=1}^{m} P(T_i \mid T_{i-1})\right] \left[\prod_{i=1}^{m} P(\ell_i \mid T_i)\right] \left[\prod_{i=1}^{m} \prod_{j=s_i}^{e_i} P(R_j \mid T_i)\right] \tag{6.9}$$

Note that (6.9) differs from (6.4) by changing the (prior) distribution of $\mathcal{S}$ from a Markov process:

$$P(\mathcal{S}) = \prod_{j=1}^{n} P(T_j \mid T_{j-1}) \tag{6.10}$$

to a *semi-Markov process*[3]:

$$P(\mathcal{S}) = \prod_{j=1}^{m} P(T_j \mid T_{j-1}) P(\ell_j \mid T_j) \tag{6.11}$$

and so (6.9) is referred to as a *hidden semi-Markov model* (HSMM). This model incorporates explicit segment length distributions conditioned on segment type.

Explicit modeling of segment length has proven useful in bioinformatics for modeling distributions of 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 4.2 (taken from (Schmidler et al., 2000)) shows data on the differing length distributions of two types of protein secondary structure segments.

[3]Note the change in indices from n to m, reflecting the change in notation: T_i goes from being the state at position i to being the state of segment i.

6.2.3 Generalized hidden Markov models and stochastic segment models

Recently, assumption (6.7) has been relaxed for applications in bioinformatics and speech recognition by allowing models of intra-segment position dependence. Such models are written in the form:

$$P(R, \mathcal{S}) \propto \prod_{j=1}^{m} P(R_{[s_j:e_j]} \mid T_j) P(T_j \mid T_{j-1}) P(\ell_j \mid T_j) \tag{6.12}$$

and have been developed 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 *segmentation* (Schmidler et al., 2000) *models*. Here we adopt the term *stochastic segment models* (SSMs) to denote the slightly more general class of models described by (6.1), and use *generalized HMMs* (GHMMs) to refer to the special case of (6.1) where the prior distribution $P(\mathcal{S})$ takes the special form (6.12).

It is perhaps worth pointing out explicitly the additional modeling capabilities obtained by adopting SSMs in place of HSMMs. SSMs relax the assumption of *intra-segment* conditional independence among sequence positions. Several models developed for protein secondary structure prediction violate this assumption, including neural networks, later versions of GOR, and other segment-based and empirical potential methods described in Chapter 3. Generally speaking, we may expand the logarithm of the joint distribution over positions in the segment:

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

$$\sum_{s \le i \le e} f_i(R_i) + \sum_{s \le (i,j) \le e} g_{i,j}(R_i, R_j) + \sum_{s \le (i,j,k) \le e} h_{i,j,k}(R_i, R_j, R_k) + \ldots \tag{6.13}$$

where it can be seen that the HSMM model, which sets all but the first set of terms on the *rhs* to zero, is quite restrictive.

Other examples of segment likelihoods which fall into the class of SSMs but not

HSMMs which are particularly relevant for modeling proteins are given in Chapter 4 and (Schmidler et al., 2000).

6.2.4 Stochastic segment interaction models

As pointed out in Section 6.1.1, SSIMs relax the assumption of conditional independence of *inter-segment* sequence positions by introducing a structured notion of segment interaction. Of course the general form of SSIMs given by (6.2) is somewhat too general, and an important tradeoff must be found in modeling strong dependencies and ignoring others.

Pair potentials

To see this generality, note that (6.2) includes standard pair potentials as a special case by setting $m = n$ the length of R and $p = 1$. If we relax the disjoint assumption of sets of interacting segments, pair potentials may be described more conveniently by taking $p = \binom{n}{2}$ all pairs of segments.

Given this, it is clear that other models for interactions in biopolymer sequences may be seen as special cases of SSIMs. For example, the models of Hubbard described in Chapter 3 impose a pair potential (3.2) model of β-sheets

RNA folding and stochastic grammars

Interestingly, RNA secondary structure prediction models based on *stochastic context-free grammars* (SCFGs) also fall into the SSIM framework. RNA secondary structure formation has been successfully modeled using statistical mechanical models and experimentally determined energetic parameters (Zuker and Stiegler, 1981; Zuker and Sankoff, 1984; Zuker, 1989; McCaskill, 1990).

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

such that $i_1 \leq i_2$, we have:

$$i_1 = i_2 \Leftrightarrow j_1 = j_2 \tag{6.14}$$

$$i_2 < j_1 \Rightarrow i_1 < i_2 < j_2 < j_1 \tag{6.15}$$

Biologically these constraints allows each nucleotide to participate in only one base pair, and restrict the structure from forming pseudo-knots. Mathematically the structure may be drawn as a planar graph, and may be represented by a context-free grammar (Searls, 1993). Stochastic context-free grammar models have been successfully applied to RNA secondary structure prediction (Eddy and Durbin, 1994; Sakakibara et al., 1994; Durbin et al., 1998).

Given such a secondary structure, the probability of a ribonucleotide sequence R may be written as a Gibbs measure:

$$P(R \mid H) \propto \exp(-U(R,H)/kT) \tag{6.16}$$

with

$$U(R,H) = \sum_{i \notin H} f(R_i) + \sum_{(i,j) \in H} g(R_i, R_j) \tag{6.17}$$

a pair potential with nucleotide-interaction free energy terms. Typically U is defined over neighboring pairs to account for effects such as base stacking.

It can be seen that the above models fall into the general form given by 6.2 above, and form a restricted subset of non-local interactions. As described in Chapter 8, this restricted class is amenable to efficient algorithms for predictive inference.

Note that it is precisely the tendency of protein β-sheets to violate the constraints given by (6.15) which makes SSIMs more desirable, and less computationally tractable. It is of interest to relate SSIMs to attempts in the RNA folding literature to account for pseudo-knots, but I do not pursue this here.

6.3 Priors on segmentations

We have not yet discussed the specification of $P(\mathcal{S})$ in SSMs. As pointed out in in Section 6.2, the definition of SSMs without interactions (6.1) adopted here (from (Schmidler et al., 2000)) is slightly more general than that of GHMMs (6.12). In particular, it allows for alternative forms for the marginal probability distribution on segmentations $P(\mathcal{S})$. From a Bayesian perspective, GHMMs are SSMs[4] with a specific form of *segmentation prior*.

It is often convenient to specify a segmentation prior by first conditioning on the the number of segments m:

$$P(\mathcal{S}) = P(m)P(\{S_j\}_{j=1}^m \mid m) \tag{6.18}$$

As noted in Section 6.2.2, implicit in GHMMs (and therefore HSMMs and standard HMMs) is an assumed prior of this form, with $P(\{S_j\}_{j=1}^m \mid m)$ factored as a semi-Markov process (6.11). However in comparing with (6.18), we can see that (6.11) embodies certain assumptions:

(i) *Uniform number segments:* We have

$$P(m) \propto 1 \tag{6.19}$$

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

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

$$P(\{T_j\}_{j=1}^m \mid m) = \prod_{j=1}^{m} P(T_j \mid T_{j-1}) \tag{6.20}$$

[4]SSMs as originally defined in (Ostendorf et al., 1996) also used this form of prior.

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

$$P(\{\ell_j\}_{j=1}^m \mid m, \{T_j\}_{j=1}^m) = \prod_{j=1}^m P(\ell_j \mid T_j) \tag{6.21}$$

The wide application of HMMs in practice suggests that these assumptions are often reasonable; however, they may be inappropriate in some domains. In this section, we describe several other possible forms of segmentation priors. In Chapter 7 we discuss how the choice of priors affects computational complexity of inference. Chapter 5 describes an application where some of these assumptions are undesirable.

6.3.1 Alternative segmentation priors

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

(i) *Uniform prior* A simple approach considers a prior that is uniform across segmentations:

$$P(\mathcal{S}) \propto 1 \tag{6.22}$$

yielding the joint distribution

$$P(R, \mathcal{S}) \propto \prod_{j=1}^m P(R_{[s_j:e_j]} \mid \mathcal{S}) \tag{6.23}$$

Again, (6.22) is improper but yields a proper posterior. Prior (6.22) is adopted implicitly in related work on DNA and protein sequence segmentation (Auger and Lawrence, 1989; Stormo and Haussler, 1994; Liu and Lawrence, 1996) and the resulting MAP segmentation (see (6.36) in Section 6.4) may be interpreted as a *maximum likelihood* segmentation of the sequence R.

(ii) *Semi-Markov process prior*: The semi-Markov process prior 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:

$$P(R, \mathcal{S}) = P(m) \prod_{j=1}^{m} P(R_{[s_j:e_j]} \mid \mathcal{S}) P(T_j \mid T_{j-1}) P(\ell_j \mid T_j) \tag{6.24}$$

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

(iii) *Sequence-specific prior*: Priors (6.22) and (6.11) model the distribution over general segmentations. We may also consider priors over segmentations of a particular sequence R, which we call *sequence-specific priors*. For example, conditioning on the observed sequence length n gives prior probability mass only for segmentations subject to the constraint[5]

$$\sum_{j=1}^{m} \ell_j = n \tag{6.25}$$

Here n becomes fixed in the model, rather than an observed quantity. This approach is used by (Liu and Lawrence, 1999), who adopt a combinatorial prior which is uniform on m-segmentations given m. Adapting this to the current formulation, we get:

$$P(\mathcal{S} \mid n) \propto P(m) \binom{n-1}{m-1}^{-1} (|\mathcal{A}_T| - 1)^{-(m-1)} / |\mathcal{A}_T| \tag{6.26}$$

Similarly, one can adapt the uniform prior (6.22) to the sequence-dependent case, where conditioning on n:

$$P(\mathcal{S} \mid n) \propto 1 \tag{6.27}$$

has no effect other than to make the prior a proper distribution.

[5]Note that under priors (6.22,6.11), the *posterior* distribution over segmentations $P(\mathcal{S} \mid R)$ has mass restricted to segmentations satisfying (6.25), but this results from the likelihood term and not the prior.

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

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

$$P(\mathcal{S}) \propto \prod_{j=1}^{m} f(e_{j-1}, e_j, T_{j-1}, T_j) \tag{6.28}$$

for some f independent of j.

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

$$P(\mathcal{S} \mid X) \propto g(X) \prod_{j=1}^{m} f(e_{j-1}, e_j, T_{j-1}, T_j \mid X) \tag{6.29}$$

We will see in Chapter 7 that segment decomposability has a dramatic impact on computational complexity of inference. Note that for a SSM given by (6.1), the posterior distribution $P(\mathcal{S} \mid R)$ is segment decomposable *if and only if* the prior $P(\mathcal{S})$ is. Priors given above are examples of this class of segment-decomposable priors.

(v) *General segmentation priors:* It is easy to construct desirable priors which do not have the decompositions specified above. For example, one may wish to specify a prior on secondary structure content via

$$P(\mathcal{S}) = P(m)P(\#(H \in T), \#(E \in T), \#(L \in T) \mid m) \tag{6.30}$$

Such general priors which depend on global features of a segmentation impose a non-decomposable structure on the joint distribution (6.1) which significantly complicates posterior inference (see Chapter 8).

As described in Chapter 7 below, the different structure imposed on the joint distribution by these various priors alters the computational complexity of algorithms for posterior inference. In Chapter 9, we will see that they may also significantly affect the quality of posterior inference.

6.3.2 Priors on segment interactions

We now consider segmentation priors in the case of SSIMs, namely the specification of $P(\mathcal{S}, \mathcal{I})$. Again, we distinguish several types:

(i) *Uniform:* Analogous to (6.22), we may adopt a uniform prior on interacting segmentations:

$$P(\mathcal{S}, \mathcal{I}) \propto 1 \tag{6.31}$$

which yields a maximum likelihood interacting segmentation.

(ii) *Conditionally uniform:* Given any of the priors discussed previously in Section 6.3.1, the prior may be extended by a *conditionally uniform* prior on segment interactions. This yields a joint prior of the form:

$$P(\mathcal{S}, \mathcal{I}) \propto P(\mathcal{S})/c(\mathcal{S}) \tag{6.32}$$

where $c(\mathcal{S})$ is the number of possible interactions formed on segments in $\mathcal{S}$, and $P(\mathcal{S})$ may for example be a semi-Markov prior of the form (6.11).

(iii) *Noninformative:* A related approach sets

$$P(\mathcal{S}, \mathcal{I}) \propto P(\mathcal{S}) \tag{6.33}$$

Note that this is equivalent to multiplying (6.32) by a factor $c(\mathcal{S})$, and so strongly favors segment types which interact, as discussed in Chapter 5.

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

$$P(\mathcal{S}, \mathcal{I}) \propto P(\mathcal{S}) P(\mathcal{I} \mid \mathcal{S}) \tag{6.34}$$

Some care is required in specifying $P(\mathcal{I} \mid \mathcal{S})$; examples of this approach are developed in Sections 5.2.1 and 5.2.3.

(v) *General informative* Alternatively, general priors on interacting segmentations may be specified. By analogy to (6.30) for example, we might write

$$P(\mathcal{S},\mathcal{I}) = P(m)P(\#(H \in T), \#(L \in T), \#(I \in \mathcal{I}) \mid m) \tag{6.35}$$

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

Sections 5.2.1 and 5.2.3 develop concrete examples of some of these forms. Experimental results are described in Chapter 9.

6.4 Inference and prediction in stochastic segment interaction models

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

6.4.1 Segmentation

We refer to the task of recovering the unobserved $(\mathcal{S},\mathcal{I})$ given an observed sequence R as *segmentation* of R. Many applications may be cast as 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.

Segmentation is inherently a problem of statistical inference, and it will not be surprising that multiple estimators of the quantities $(\mathcal{S},\mathcal{I})$ exist which exhibit different properties[6].

We first consider the case of SSMs without interactions. The most commonly used estimate of $\mathcal{S}$ is the *maximum a posteriori* (MAP) value, which we refer to as the

[6]The *computational* problem of segmentation concerns the development of efficient algorithms for calculating various estimators, and is discussed in Chapters 7 and 8.

MAP segmentation:

$$\begin{aligned} \mathcal{S}_{MAP} &= \arg\max_{\mathcal{S}} P(\mathcal{S} \mid R, \theta) \\ &= \arg\max_{(m, \{\ell_i\}_{i=1}^m, \{T_i\}_{i=1}^m)} P(m, (\ell_i, T_i)_{i=1}^m \mid R, \theta) \end{aligned} \tag{6.36}$$

where θ denotes the model parameters. If we define the *expected loss* of a segmentation estimator $\mathcal{S}^*$:

$$E_{\mathcal{S}|R,\theta}(L, \mathcal{S}^*) = \sum_{\mathcal{S}} P(\mathcal{S} \mid R, \theta) L(\mathcal{S}, \mathcal{S}^*) \tag{6.37}$$

for some loss function L, it is easily checked that the MAP segmentation (6.36) minimizes the expected loss under *0-1 loss*:

$$L(\mathcal{S}, \mathcal{S}^*) = \begin{cases} 0 & \text{for } \mathcal{S} = \mathcal{S}^* \\ 1 & \text{otherwise} \end{cases} \tag{6.38}$$

and is therefore an optimal estimator of $\mathcal{S}$ in the sense of minimizing Bayesian risk (Berger, 1985).

An alternative estimator for $\mathcal{S}$ is the *marginal mode* predictor, often referred to as "smoothing" in the case of HMMs:

$$\mathcal{S}_{Mode} = \{\arg\max_{T} P(T_{R_{[i]}} \mid R, \theta)\}_{i=1}^n \tag{6.39}$$

where $P(T_{R_{[i]}} \mid R, \theta)$ denotes the marginal posterior distribution over segment types at a single position i in the sequence:

$$P(T_{R_{[i]}} \mid R, \theta) = \sum_{(m,S,T)} P(m, S, T \mid R, \theta) \mathbf{1}_{\{T_{R_i} = t\}} \tag{6.40}$$

Hence (6.39) provides the sequence of marginal posterior modes at each position. Note that (6.39) involves marginalization over all possible segmentations.

It is easily checked that the marginal mode predictor (6.39) is optimal under a

Hamming distance loss function:

$$L(\mathcal{S}, \mathcal{S}^*) = \sum_{i=1}^{n} \mathbf{1}_{\{T_{R_i} = T^*_{R_i}\}} \tag{6.41}$$

which counts the number of positions assigned to an incorrect segment type. If our interest is in maximizing the number of correctly classified positions, the marginal mode predictor (6.39) is to be preferred. Chapter 9 reports experiments showing a significant improvement of $\mathcal{S}_{Mode}$ over $\mathcal{S}_{MAP}$ when applied to protein secondary structure prediction. A disadvantage of $\mathcal{S}_{Mode}$ is that the resulting segmentations may have posterior mass zero.

6.4.2 Segmentation with interactions

Segmentation under the more general framework of SSIMs introduced in this dissertation proceeds in an analogous manner. In a direct parallel to predictors (6.36,6.39), we define

$$(\mathcal{S}, \mathcal{I})_{MAP} = \arg\max_{(\mathcal{S},\mathcal{I})} P(\mathcal{S}, \mathcal{I} \mid R) \tag{6.42}$$

$$\mathcal{S}^{\mathcal{I}}_{Mode} = \{\arg\max_{T} P(T_{R_{[i]}} \mid R)\}_{i=1}^{n} \tag{6.43}$$

where $P(T_{R_{[i]}} \mid R)$ again denotes the marginal posterior distribution over structural types at a single position i in the sequence, but now marginalized over the significantly enlarged space of segmentations *including segment interactions*:

$$P(T_{R_{[i]}} \mid R) = \sum_{(\mathcal{S},\mathcal{I})} P(\mathcal{S}, \mathcal{I} \mid R) \mathbf{1}_{\{T_{R_i} = t\}} \tag{6.44}$$

As before, MAP and $Mode$ predictors (6.42,6.43) are Bayes estimates under *0-1* and Hamming distance losses, respectively. However, as in Section 5.2.3 of Chapter 5, the space of possible segmentations has been considerably expanded via the introduction of segment interaction parameters, and the same discussion of interaction priors is relevant. Since (6.31) yields a marginal prior $P(\mathcal{S})$ which is highly biased towards

interacting segments, it is expected to perform very poorly in prediction of marginal quantities $P(\mathcal{S})$ via (6.43). On the other hand, prior (6.32)) significantly downweights the effects of interaction for any particular interaction, and so is unlikely to improve marginal predictions through incorporation of joint-segment information. Instead, we may focus on prediction of the interactions themselves.

6.4.3 Contact map prediction

It should be noted that the *maximum a posteriori* segmentation $(\mathcal{S},\mathcal{I})_{MAP}$ defined by (6.42) has a different form than that of $\mathcal{S}_{MAP}$ in the SSM case (6.36). This difference is due to the change in definition: specification of a segmentation now requires specification of the interaction parameters defined in Section 6.1.1 as well. Hence the MAP segmentation is no longer a set $\mathcal{S}$, but a set $(\mathcal{S},\mathcal{I})$ where $\mathcal{I}$ as previously defined represents the sets of segment interactions and their associated parameters. In other words, $(\mathcal{S},\mathcal{I})_{MAP}$ provides the *maximum a posteriori* set of segment interactions, in addition to the segment locations and types. In the context of the β-sheet models developed in Chapter 5, this means the MAP values of sheet parameters p, $\{\mathcal{H}\}_{j=1}^{p}$, and $\{h_{i,j}\}_{i,j=1}^{p}$ are specified, identifying the strands involved in sheets and their topology and relative orientations. Hence the $Struct_{MAP}$ predictor includes β-sheet assignment and topology in addition to secondary structure assignment.

In the case of SSIMs, it is also of interest to look specifically at predicted interactions. As in the example of β-sheet prediction developed in Chapter 5, we may summarize interactions by a *contact map*, a matrix $C_{n\times n} = \{c_{i,j}\}$ where

$$c_{i,j} = \begin{cases} 1 & \text{If } R_i \text{ and } R_j \text{ are paired in a segment interaction} \\ 0 & \text{otherwise} \end{cases} \tag{6.45}$$

We will denote the contact map derived from $(\mathcal{S},\mathcal{I})_{MAP}$ as C_{MAP}:

$$C_{MAP} = \{c_{i,j}\}_{i,j=1}^{n} = \begin{cases} 1 & R_i,\, R_j \text{ paired in } (\mathcal{S},\mathcal{I})_{MAP} \\ 0 & \text{otherwise} \end{cases} \tag{6.46}$$

C_{MAP} therefore represents an estimator of C using the single highest probability set of interactions, as found in $(\mathcal{S},\mathcal{I})_{MAP}$. However we have already seen reasons that MAP estimation is undesirable. Unfortunately, the marginalization involved in computing $\mathcal{S}^{\mathcal{I}}_{Mode}$ via (6.44) at each position fails to retain the interactions of any particular segmentation, and so unlike $(\mathcal{S},\mathcal{I})_{MAP}$, the predictor $\mathcal{S}^{\mathcal{I}}_{Mode}$ does not predict the topology of interactions. However, we may also estimate C in an analogous manner to (6.44), by defining the *marginal predicted contact map*, given by the matrix:

$$C_{Mode} = \{c_{i,j}\}_{i,j=1}^{n}$$

where

$$c_{i,j} = P(i \leftrightarrow j) = \sum_{\mathcal{S},\mathcal{I}} P(\mathcal{S},\mathcal{I} \mid R)\mathbf{1}_{\{i \leftrightarrow j\}} \tag{6.48}$$

with $P(\mathcal{S},\mathcal{I})$ given by (6.2). As with $Struct_{Mode}$, the matrix C_{Mode} predicts each potential contact marginalized over *all possible segmentations and segment interactions.* Hence it may be expected to provide more accurate predictions of C than C_{MAP}, as it yields a Bayes estimator for C under a Hamming distance loss.

Calculation of the quantities $(\mathcal{S},\mathcal{I})_{MAP}$, $\mathcal{S}^{\mathcal{I}}_{Mode}$, C_{MAP}, and C_{Mode} defined here for SSIMs is more difficult than the analogous computations in SSMs, and will be discussed in Chapter 8. Experiments with using the matrices C_{MAP} and C_{Mode} as predictors of true contacts in the context of protein structure is described in Chapter 9.

Chapter 7

Dynamic Programming Algorithms for Stochastic Segment Models

In this Chapter I discuss computation with the SSM models described in Chapter 6. Efficient algorithms are given for several types of SSMs using dynamic programming. The next chapter will introduce computational methods for inference with the more general class of SSIMs introduced in Chapter 6.

7.1 Algorithms for inference in stochastic segment models

In order to perform prediction and inference with SSMs, we require algorithms for computing quantities such as (6.36) and (6.39) given an observed sequence R. As mentioned briefly in Section 6.3, the algorithms required depend on the form of the joint distribution (6.1), including the form of the prior. We consider three cases: segment-decomposable priors; *conditionally* segment-decomposable priors; and non-decomposable priors. See Section 6.3.1 for definitions.

7.1.1 Segment-decomposable priors

The most efficient algorithms for SSMs (barring special cases such as HMMs) occur for segment-decomposable priors. Several examples of segment-decomposable priors are given in Chapter 6. For example:

(i) Uniform priors (6.22 and 6.27)

(ii) Semi-Markov process priors (6.24) with $P(m) \propto 1$ improper (6.11)

In this case, the joint distribution has conditional independence structure similar to that of a hidden semi-Markov model (6.9). Thus computation can be done exactly using a generalization of the standard forward-backward and Viterbi algorithms for hidden Markov models (HMMs) to the case of hidden semi-Markov models (HSMMs), as described in (Rabiner, 1989; Stormo and Haussler, 1994; Schmidler et al., 2000). In particular, the MAP segmentation (6.36) may be calculated using the forward variables:

$$\delta(j,t) = \max_{\binom{v=1,\ldots,j-1}{l \in \mathcal{A}_T}} [\delta(v,l) f(e_- = v, e = j, T_- = l, T = t \mid \theta)] \tag{7.1}$$

with appropriate initialization, in a procedure analogous to the Viterbi algorithm for HMMs (Rabiner, 1989). Here $e_{+/-}$ represents the endpoint of the next/previous segment.

For example, the SSM for protein secondary structure prediction developed in Chapter 4 has a prior of the form (6.11), and the above becomes:

$$\delta(j,t) = \max_{\binom{v=1,\ldots,j-1}{l \in \mathcal{A}_T}} \left[\begin{array}{ll} \delta(v,l) & P(R_{[v+1:j]} \mid e_- = v, e = j, T = t, \theta) \ \times \\ & P(e = j \mid T = t, e_- = v, \theta) P(T = t \mid T_- = l, \theta) \end{array} \right] \tag{7.2}$$

The algorithm proceeds by recursively calculating $\delta(j,t)$ for $j = 1, \ldots, n$ and $t \in \mathcal{A}_T$, and then reconstructing the MAP segmentation by setting

$$\begin{aligned} S_m^* &= n \\ T_m^* &= \arg\max_{l \in \mathcal{A}_T} \delta(n,l) \end{aligned}$$

and tracing backwards.

This calculation requires $O(n^3)$ steps. Often, a maximum segment length D may be imposed, in which case the maximization (7.1) ranges over $v = j-D, \ldots, j-1$ and the algorithm is linear ($O(nD^2)$) in n. Experiments in Chapter 9 set $D = 30$, large enough to account for nearly all observed structural segments (Figure 4.2). The model given by (4.4) in fact allows reduction to $O(nD)$, but this does not hold in general, and the additional computational savings is unnecessary for protein sequences.

The recursion (7.1) provides an efficient algorithm for computing the MAP segmentation (6.36). The marginal posterior distributions $P(T_{R_{[i]}} \mid R)$ required for (6.39) may be calculated similarly, by defining forward variables:

$$\alpha(j,t) = \sum_{v=1}^{j-1} \sum_{l \in \mathcal{A}_T} \alpha(v,l) f(e_- = v, e = j, T_- = l, T = t \mid \theta) \tag{7.3}$$

and backward variables:

$$\beta(j,t) = \sum_{v=j+1}^{n} \sum_{l \in \mathcal{A}_T} \beta(v,l) f(e = j, e_+ = v, T = t, T_+ = l \mid \theta) \tag{7.4}$$

For the protein secondary structure models of Chapter 4, these become:

$$\alpha(j,t) = \sum_{v=1}^{j-1} \sum_{l \in \mathcal{A}_T} \alpha(v,l) P(R_{[v+1:j]} \mid e_- = v, e = j, T = t, \theta) \times P(e = j \mid T = t, e_- = v, \theta) P(T = t \mid T_- = l, \theta) \tag{7.5}$$

and

$$\beta(j,t) = \sum_{v=j+1}^{n} \sum_{l \in \mathcal{A}_T} \beta(v,l) P(R_{[j+1:v]} \mid e_+ = v, e = j, T_+ = l, \theta) \times P(e_+ = v \mid e = j, T_+ = l, \theta) P(T_+ = l \mid T = t, \theta) \tag{7.6}$$

Given (7.3,7.4), we may then compute the marginal posteriors (6.40) required for (6.39) via:

$$P(T_{R_i} = t \mid R, \theta) = \sum_{j=i-D+1}^{i-1} \sum_{k=i}^{j+D-1} \sum_{l \in \mathcal{A}_T} \alpha(j,l)\beta(k,t) f(e_- = j, e = k, T_- = l, T = t\theta)/Z \quad (7.7)$$

using (7.3,7.4) above. Here Z is the normalizing constant or marginal likelihood $P(R \mid \theta)$, available directly from the forward pass (7.3). Calculation of (7.3) and (7.4) can be done in $O(n^3)$ (or $O(nD^2)$ for fixed D), and (7.7) yields the marginal posterior distribution at each position in $O(n^2)$ (or $O(nD)$) time, using the following calculation:

$$\begin{aligned} P(T_{R_{[i+1]}} = t \mid R, \theta) = & \quad (7.8) \\ P(T_{R_{[i]}} = t) \quad & - \sum_{j=i-D}^{i-1} \sum_{l \in \mathcal{A}_T} \left[\alpha(j,l)\beta(i,t) f(e_- = j, e = i, T_- = l, T = t \mid \theta)\right]/Z \\ & + \sum_{k=i+1}^{i+D} \sum_{l \in \mathcal{A}_T} \left[\alpha(i,l)\beta(i+1,k) f(e_- = i, e = k, T_- = l, T = t \mid \theta)\right]/Z \end{aligned}$$

In the context of the protein models of Chapter 4, (7.7) becomes:

$$\begin{aligned} P(T_{R_i} = t \mid R, \theta) = & \sum_{j=i-D+1}^{i-1} \sum_{k=i}^{j+D-1} \sum_{l \in \mathcal{A}_T} \alpha(j,l)\beta(k,t) P(R_{[j+1:k]} \mid e_- = j, e = k, T = t, \theta) \times \\ & P(e = k \mid e_- = j, T = t, \theta) P(T = t \mid T_- = l, \theta)/Z \quad (7.9) \end{aligned}$$

and (7.8) becomes

$$\begin{aligned} P(T_{R_{[i+1]}} = t \mid R, \theta) = \\ P(T_{R_{[i]}} = t) \quad - \sum_{j=i-D}^{i-1} \sum_{l \in \mathcal{A}_T} \big[\alpha(j,l)\beta(i,t)P(R_{[j+1:i]} \mid e_- = j, e = i, T = t, \theta) \\ P(e = i \mid e_- = j, T = t, \theta)P(T = t \mid T_- = l, \theta)\big]/Z \\ + \sum_{k=i+1}^{i+D} \sum_{l \in \mathcal{A}_T} \big[\alpha(i,l)\beta(i+1,k)P(R_{[i+1:k]} \mid e_- = i, e = k, T = t, \theta) \\ P(e = k \mid e_- = i, T = t, \theta)P(T = t \mid T_- = l, \theta)\big]/Z \end{aligned} \tag{7.10}$$

The algorithms given in this section demonstrate that both predictors (6.36,6.39) may be calculated efficiently under segment-decomposable priors.

7.1.2 Conditionally segment-decomposable priors

I consider here the specific case where the prior $P(\mathcal{S})$ is conditionally segment-decomposable given m the number of segments. Thus the prior is of the form

$$P(\mathcal{S}) \propto g(m \mid \theta) \prod_{j=1}^{m} f_j(e_{j-1}, e_j, T_{j-1}, T_j \mid m, \theta) \tag{7.11}$$

with $g(m, \theta)$ itself not segment-decomposable. Examples include:

(i) Semi-Markov process priors (6.24) with non-uniform marginal prior $P(m)$, where

$$\begin{aligned} g(m \mid \theta) &= P(m) \\ &\text{and} \\ f_i(e_-, e, T_-, T \mid m, \theta) &\equiv f(e_-, e, T_-, T \mid m, \theta) \\ &= P(e = j \mid T = t, e_- = v, \theta)P(T = t \mid T_- = l, \theta) \end{aligned}$$

(ii) Sequence dependent priors with $P(\mathcal{S} \mid n) = h(m, n)$ such as (6.26), where

$$\begin{aligned} g(m \mid n, \theta) &= h(m, n) \\ &\text{and} \\ f_i(S_-, S, T_-, T \mid m, n, \theta) &\equiv 1 \end{aligned}$$

In this case, the algorithms of the previous section do not apply, and must be adapted. Algorithms for this case are given by (Auger and Lawrence, 1989; Liu and Lawrence, 1996) utilizing the following forward and backward variables:

$$\delta(j,t,k) = \max_{\binom{v=1,\ldots,j-1}{l\in\mathcal{A}_T}} \big[\delta(v,l,k-1)P(R_{[v+1:j]} \mid e_{k-1}=v, e_k=j, T_k=t,\theta)\times f_k(e_{k-1}=v, e_k=j, T_{k-1}=l, T_k=t,\theta)\big] \quad (7.12)$$

$$\alpha(j,t,k) = \sum_{v=1}^{j-1}\sum_{l\in\mathcal{A}_T} \alpha(v,l,k-1)P(R_{[v+1:j]} \mid e_{k-1}=v, e_k=j, T_k=t,\theta)\times f_k(e_{k-1}=v, e_k=j, T_{k-1}=l, T_k=t,\theta) \quad (7.13)$$

$$\beta(j,t,k) = \sum_{v=j+1}^{n}\sum_{l\in\mathcal{A}_T} \beta(v,l,k-1)P(R_{[j+1:v]} \mid e_{+}=v, e=j, T_{+}=l,\theta)\times f_k(e_{=}j, e_{+}=v, T=t, T_{+}=l,\theta) \quad (7.14)$$

defined for $k = 1, \ldots, n$.

As before, the MAP segmentation can be reconstructed by setting:

$$\begin{aligned} (m*, T^*_{m^*}) &= \arg \max_{\binom{m\in\{1,\ldots,n\}}{l\in\mathcal{A}_T}} \delta(n,l,m)g(m) \quad (7.15)\\ S^*_{m^*} &= n \\ T^*_{m^*} &= \arg\max_{l\in\mathcal{A}_T} \delta(n,l,m^*) \end{aligned}$$

and tracing back recursively. (Note that the MAP segmentation does not necessarily have the MAP number of segments m.) Here the forward/backward computations require $O(n^4)$ (or $O(n^2D^2)$) operations, a factor of n more expensive than those in the previous section.

In this case, computation of the marginal posterior distributions $P(T_{R_{[i]}} \mid R)$ requires marginalization over m:

$$P(T_{R_i} = t \mid R, \theta) = \sum_{m=1}^{n} g(m \mid \theta) \sum_{q=1}^{m} \sum_{j=i-D+1}^{i-1} \sum_{k=i}^{j+D-1} \sum_{l \in \mathcal{A}_T} \alpha(j, l, q-1)\beta(k, t, m-q+1) \times f_q(e_{q-1} = j, e_q = k, T_{q-1} = l, T_q = t \mid \theta)/Z \quad (7.16)$$

The above provide efficient algorithms for calculation of posterior quantities such as $\mathcal{S}_{MAP}$ and $\mathcal{S}_{Mode}$ in the case of SSMs with conditionally segment-decomposable priors $P(\mathcal{S})$.

7.1.3 Non-decomposable priors

In the case of general non-decomposable priors of the form (6.30), efficient algorithms do not exist. However computations may be done approximately using the techniques introduced in the next Chapter of this dissertation for inference with SSIM models.

7.2 General remarks

The previous sections provided efficient algorithms for calculation of the $\mathcal{S}_{MAP}$ and $\mathcal{S}_{Mode}$ predictors of secondary structure under several classes of models. It is worth reiterating that (7.7,7.16) provide the *exact* marginal posterior distribution over segment types at each position, averaging over *all possible* segmentations, and hence provide an exact measure of the uncertainty of prediction at each position (subject to modeling assumptions). Figure 9.2 in Chapter 9 shows that this measure correlates very strongly with prediction accuracy. Figure 9.1 shows how this uncertainty varies

along true and predicted segments.

Chapter 8

Markov Chain Monte Carlo Algorithms for Stochastic Segment Interaction Models

Previous chapters introduced the formalism of segment interactions and SSIMs (6.2), and showed an example of their use in modeling protein β-sheets. In this chapter I discuss computational issues which arise in dealing with SSIMs. Unlike the case of SSMs discussed in Chapter 7, efficient algorithms do not exist for SSIMs beyond certain special cases. After discussing one such case, I focus instead on developing Markov chain Monte Carlo simulation algorithms for approximate inference and prediction with general SSIMs.

8.1 Computing with stochastic segment interaction models

The joint distribution given by (6.2) is easily evaluated for any fixed segmentation $(\mathcal{S}, \mathcal{I})$ of a sequence R. However, serious difficulties arise in attempting to calculate posterior quantities such as $(\mathcal{S}, \mathcal{I})_{MAP}$, $\mathcal{S}^{\mathcal{I}}_{Mode}$, C_{MAP}, or C_{Mode}, or the normalizing constant $Z = P(R)$ in SSIMs. This is because the conditional independence structure

of SSMs (6.1) is critical for recursive decomposition of the joint distribution $P(R, \mathcal{S})$, which in turn enables the efficient exact calculation of posterior probabilities via dynamic programming algorithms described in Chapter 7. Although I will show that restricted classes of SSIMs may still admit efficient algorithms, general SSIMs will require other solutions. To address this difficulty, I develop a set of Markov chain Monte Carlo algorithms for approximate computation of posterior quantities under SSIM models.

8.2 Exact calculation with limited interactions

Before treating general SSIMs, I consider limited segment interactions such as those introduced in Chapter 5 for modeling β-hairpins. We will see that inference in such models may be performed exactly via dynamic programming. This is useful both for prediction with such restricted models, and for understanding the modeling issues of SSIMs without the confounding issue of approximate computation.

The key assumption of the β-hairpin model defined in Section 5.2.1 is that a β-hairpin consists of two strand segments separated by a single loop. This restricts the maximum distance between interacting segments, but more importantly it prevents "interleaving" of segments in different β-hairpins, which violate the constraints of (6.15). Because of this restriction on the interaction parameters, exact calculations may be done in polynomial time. (This should not be surprising, as we may consider the strand-loop-strand structure of a β-hairpin as a "supersegment" of sorts, and apply the algorithms of Chapter 7, modified slightly to account for strand interaction parameters within this supersegment.) Although I consider only β-hairpin models here, it is clear that stochastic context-free grammar (SCFG) algorithms used for models of the form (6.16) may be applied here, and this is an interesting area for further study.

One way to perform the exact calculation with β-hairpin models is by defining

the forward variables in terms of segment triples, for example using:

$$\delta(j,t,j_{-1},t_{-1},j_{-2},t_{-2}) = \max_{\substack{v=1,\ldots,j_{-2}-1 \\ l\in\mathcal{A}_T}} P(R_{[v+1:j]} \mid S=j, S_{-1}=j_{-1}, S_{-2}=j_{-2}, S_{-3}=v, T=t, T_{-1}=t_{-1}, T_{-2}=t_{-2}, T_{-3}=l, \theta) \times priors \quad (8.1)$$

and the analogous definitions of α and β forward and backward variables for computation of marginal distributions. Again, more efficient calculations may be possible using SCFG-type algorithms, and this will be explored in future work.

8.3 Markov chain Monte Carlo segmentation

Unfortunately, more general forms of SSIMs do not lend themselves to such efficient algorithms. In general, introduction of joint-segment models into SSMs makes exact calculation of posterior probabilities intractable. Nevertheless, approximate inference in these models may be achieved using Monte Carlo approximation. While it appears quite difficult to develop approximation schemes that are provably efficient in a formal sense (Motwani and Raghavan, 1995), I show that Monte Carlo inference based on Markov chain simulation provides satisfactory empirical performance.

Markov chain Monte Carlo (MCMC) inference has become a standard tool in the Bayesian statistics community for inference with complex models (Gelfand and Smith, 1990; Smith and Roberts, 1993; Gelman et al., 1995; Gilks et al., 1996). In this Section I introduce basic concepts of MCMC required to develop an algorithm for inference with models of the form (6.2). For the purposes of this dissertation, it is sufficient to introduce two MCMC approaches, the *Metropolis-Hastings algorithm* and the *Gibbs sampler*.

8.3.1 Metropolis-Hastings and Gibbs sampling

The primary tool to be used in developing a MCMC segmentation algorithm will be the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970).

Let X be a finite set, and $\pi(x)$ for $x \in X$ a probability distribution on X. The Metropolis-Hastings construction yields a Markov chain on X with transition kernel defined by the product of a proposal distribution $T(x, y)$, $x, y \in X$ and an acceptance ratio:

$$\rho(x, y) = \min \left\{ 1, \frac{\pi(y)T(y, x)}{\pi(x)T(x, y)} \right\} \tag{8.2}$$

The algorithm proceeds by iterating the following steps:

Algorithm 8.1 (Metropolis-Hastings).

1) Initialize $x^{(0)}$

2) Iterate:

(a) For state $x^{(t)}$ *at time* t*, draw* $y \sim T(y \mid x^{(t)})$.

(b) Set

$$x^{(t+1)} = \begin{cases} y & \textit{with probability } \rho(x, y) \\ x^{(t)} & \textit{otherwise} \end{cases}$$

It is easily checked that the resulting transition kernel is reversible with respect to π:

$$\pi(x)P(x, y) = \pi(y)P(y, x) \quad \forall x, y \in X \tag{8.4}$$

where $P(x, y) = T(x, y)\rho(x, y)$, and therefore P has unique stationary measure π for P ergodic. Therefore Algorithm 8.1 defines a Markov chain with stationary distribution $\pi(x)$, and simulation yields (dependent) samples from $\pi(x)$.

It is also useful to briefly describe a special case of the Metropolis-Hastings (MH) algorithm known as *Gibbs sampling* (Geman and Geman, 1984; Smith and Roberts, 1993). For a multidimensional state space $X = X_1 \times X_2 \times \ldots \times X_k$, and $X \ni x = (x_1, x_2, \ldots, x_k)$, the Gibbs sampler iteratively samples from the conditional distributions:

Algorithm 8.2 (Gibbs Sampling).

1) Initialize $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_k^{(0)})$

2) Iterate:
for $i = 1 : k$ *draw* $x_i^{(t+1)} \sim \pi(x_i \mid x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_k^{(t)})$

Algorithm 8.2 is a *systematic-scan* Gibbs sampler; a random scan may also be used. The Gibbs sampler requires direct sampling from conditional distributions $\pi(x_i \mid x_{\{j \neq i\}})$, which is often infeasible. Note that exact conditional sampling yields Metropolis-Hastings moves with acceptance probability $\rho(x, y) = 1$. Hence we may refer to both schemes as MH samplers when convenient.

8.3.2 Reversible-jump Markov chain Monte Carlo segmentation

The MCMC approach may be applied to inference under posterior distribution $P(\mathcal{S}, \mathcal{I} \mid R)$ for models (6.2), by taking $\pi = P(\mathcal{S}, \mathcal{I} \mid R) \propto (6.2)$. Because the dimension of the parameter space varies during the Markov chain simulation (for example, the number of segments m and the segment interactions $\mathcal{I}$ are among the parameters being inferred), the Metropolis-Hastings scheme is applied using a reversible-jump approach (Green, 1995). Briefly, this requires that jumps between models of differing dimension also be reversible.

Markov chain Monte Carlo segmentation with independent segment models

To begin, I describe a reversible-jump MCMC algorithm for sampling from SSM joint distributions (6.1). Although exact algorithms exist for SSMs as seen in Chapter 7, this MCMC algorithm will be extensible to models of form (6.2). The construction of a Markov chain on the space of segmentations will use the following set of Metropolis proposals:

- *Type change*:
  Given a set of segments $\mathcal{S} = (m, S, T) = (S_1, \ldots, S_m, T_1, \ldots, T_m)$, propose a

move to segments $\mathcal{S}^* = (m, S, T^*)$ with $T^* = (T_1, \ldots, T_{k-1}, T_k^*, T_{k+1}, \ldots, T_m)$, where $T_k^* \sim Uniform[\{H, E, L\}]$ for k chosen uniformly at random or via systematic scan.

- *Position change*:
  Given $\mathcal{S}$, propose $\mathcal{S}^* = (m, S^*, T)$ with $S^* = (S_1, \ldots, S_{k-1}, S_k^*, S_{k+1}, \ldots, S_m)$, where $S_k^* \sim Uniform[S_{k-1} + 1, S_{k+1} - 1]$.

- *Segment split*:
  Given $\mathcal{S}$, propose $\mathcal{S}^* = (m^*, S^*, T^*)$ with $m^* = m + 1$ by splitting segment k into two new segments $(k^*, k^* + 1)$ as follows:

  (i) Set $S_{k^*+1} = S_k$

  (ii) Set $S_{k^*} \sim Uniform[S_{k-1} + 1, S_k - 1]$

  (iii) With probability $\frac{1}{2}$, set $T_{k^*} = T_k$ and $T_{k^*+1} = T_{new}$ with $T_{new} \sim Uniform[\{H, E, L\}]$; 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 $m^* = m - 1$.

Type change and *Position change* moves may be performed directly by Gibbs sampling, and hence require no acceptance criteria ($\rho = 1$). (It may still be more efficient to *Metropolize* such moves (Liu, 1996).) The form of (6.1) makes exact calculation of conditionals efficient, involving only terms that are *local* with respect to the affected segment:

$$P(T_k = t \mid \mathcal{S} \setminus \{T_k\}) \propto P(T_k = t \mid T_{k-1}) P(S_k \mid S_{k-1}, T_k = t) P(T_{k+1} \mid T_k = t) \times P(R_{[S_{k-1}+1:S_k]} \mid S_{k-1}, S_k, T_k = t)$$

Segment split and *Segment merge* moves jump between models of different dimension, and are accepted or rejected according to a reversible-jump Metropolis criteria. For

example, for *Segment split* the acceptance probability becomes:

$$\rho_{split(k)}(\mathcal{S}, \mathcal{S}^*) = \frac{\prod_{j=1}^{m+1} P(T_j^* \mid T_{j-1}^*) P(S_j^* \mid S_{j-1}^*, T_j^*) P(R_{[S_{j-1}^*+1:S_j^*]} \mid S_{j-1}^*, S_j^*, T_j^*)}{\prod_{j=1}^{m} P(T_j \mid T_{j-1}) P(S_j \mid S_{j-1}, T_j) P(R_{[S_{j-1}+1:S_j]} \mid S_{j-1}, S_j, T_j)} \times \frac{m(S_k - S_{k-1} - 1)|\mathcal{A}_T|}{(m+1)} \quad (8.5)$$

Again, the factorization of (6.1) allows this ratio to be calculated *locally* when dimension changes move to subsets or supersets of the current segmentation, which holds for the *Segment split*, *Segment merge* moves defined above.

$$\rho(\mathcal{S}, \mathcal{S}^*) = \frac{\prod_{j=k}^{k+1} P(R_{[S_{j-1}^*+1:S_j^*]} \mid S_{j-1}^*, S_j^*, T_j^*) P(S_j^* \mid S_{j-1}^*, T_j^*) P(T_{j+1}^* \mid T_j^*)}{P(R_{[S_{j-1}+1:S_j]} \mid S_{j-1}, S_j, T_j) P(S_j \mid S_{j-1}, T_j) P(T_{j+1} \mid T_j)} \times \frac{m(S_k - S_{k-1} - 1)|\mathcal{A}_T|}{(m+1)} \quad (8.6)$$

Hence the full joint distribution (6.1) need not be evaluated at each step.

Combining these four steps yields the following algorithm for MCMC segmentation of a protein sequence under the joint distribution (6.1):

Algorithm 8.3 (MCMC Segmentation - Independent segments).

1) Initialize $\mathcal{S}^{(0)} = (m^{(0)}, S^{(0)}, T^{(0)})$.

2) Iterate:

(a) Draw $k \sim Uniform[1, m^{(t)}]$

(b) Set

$$\mathcal{S}' = \begin{cases} segment\, split(k) & with\ probability\ .5 \\ segment\, merge(k) & otherwise \end{cases}$$

(c) Set

$$\mathcal{S}^{(t+1)} = \begin{cases} (m', S', T') & \textit{with probability } \rho_{split(k)} \\ (m^{(t)}, S^{(t)}, T^{(t)}) & \textit{otherwise} \end{cases}$$

(d) For $k = 1 : m^{(t+1)}$
draw $S_k^{(t+1)} \sim P(S_k \mid m^{(t+1)}, S_{j<k}^{(t+1)}, S_{j>k}^{(t)}, T_{j<k}^{(t+1)}, T_{j\geq k}^{(t)})$

(e) For $k = 1 : m^{(t+1)}$
draw $T_k^{(t+1)} \sim P(T_k \mid m^{(t+1)}, S_{j\leq k}^{(t+1)}, S_{j>k}^{(t)}, T_{j<k}^{(t+1)}, T_{j>k}^{(t)})$

In general, the parameters $(S_j^{(t)}, T_j^{(t)})$ are dependent and so it is more efficient to draw them jointly (Liu et al., 1994). However this would significantly complicate things when the algorithm is extended to consider segment pairing during the sampling of T_j's as well.

The following result for the algorithm described can now be stated:

Lemma 8.1. *The Markov chain constructed by Algorithm 8.3 has invariant distribution given by (6.1).*

Proof: Each individual move is invariant with respect to (6.1) by construction, and both systematic and random scans of reversible moves are also reversible (see for example (Gelman et al., 1995)). ◇

Together the 4 steps given above suffice to obtain an ergodic Markov chain. However, it is helpful to add two additional moves which facilitate mixing of the Markov chain:

- *Segment introduction*:
  Given $\mathcal{S}$, propose $\mathcal{S}^*$ with $m^* = m + 2$ segments by splitting segment k into three segments $(k, k+1, k+2)$:

  (i) Draw $k \sim Uniform[1, m]$

  (ii) Set $S_{k+2}^* = S_k$

  (iii) Draw $l_1 \sim Uniform\,[S_{k-1} + 1, S_k - 1]$

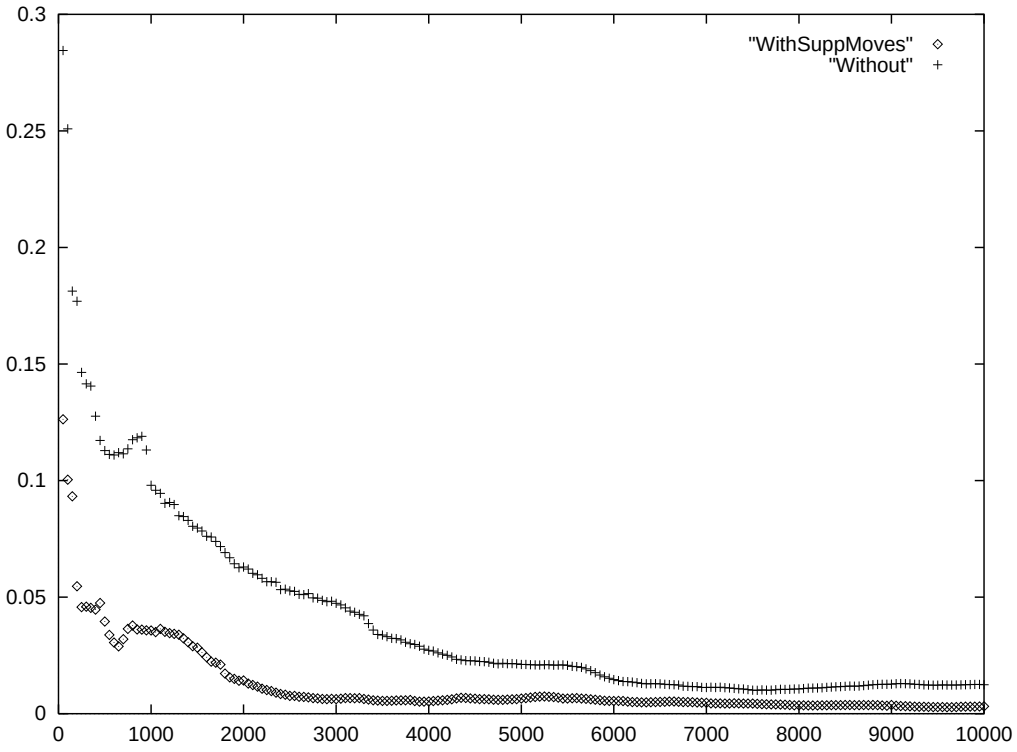


Figure 8.1: Convergence of MCMC sampling algorithm. Plot shows mean Kullback-Leibler (KL) divergence between marginal distributions $P(T_{R_{[i]}})$ obtained from exact and MCMC calculations for sequence 5nul using two sampling algorithms - with and without supplemental Metropolis moves. KL divergence between two probability distributions $\mathbf{p}$ and $\mathbf{q}$ is defined as $KL(\mathbf{p}, \mathbf{q}) = \sum_i p_i \log(\frac{p_i}{q_i})$.

(iv) Draw $l_2 \sim Uniform\,[S_{k-1}+1, S_k-1] \setminus \{l_1\}$

(v) Set $S_k^* = \min(l_1, l_2)$

(vi) Set $S_{k+1}^* = \max(l_1, l_2)$

(vii) Set $T_k^* = T_{k+2}^* = T_k$

(viii) Draw $T_{k+1}^* \sim Uniform[\mathcal{A}_T \backslash \{T_k\}]$.

- *Segment removal*:
  Similar to *segment introduction*, but segment k with $T_{k+1} = T_{k-1}$ is removed and merged with its immediate neighbors, yielding $m^* = m - 2$.

These are dimension-altering Metropolis moves, and calculation of the reversible-jump Metropolis acceptance follows exactly as above. The result is an increase in the mixing rate of the underlying Markov chain (see Figure 8.1).

MCMC segmentation with joint-segment models

In order to perform inference under joint segment models (6.2), the Metropolis moves above must be supplemented by additional moves involving interacting segments:

- *Segment join*:
  Given an interacting segmentation $(\mathcal{S}, \mathcal{I})$, propose a new interaction I^{new} involving two non-interacting segments S_j and S_k, so that $I^* = I \cup \{I\}$. The proposal must also generate interaction parameters H^{new}, and the acceptance ratio must account for this. For the β-sheet application described in Chapter 5, this move takes two non-interacting β-strands and proposes to join them into a β-sheet, generating the register at random.

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

- *Segment align*:
  Given a segment interaction $I \in \mathcal{I}$, sample the associated interaction parameters. For β-sheets, samples the register of paired β-strands.

- *Segment insert* and *Segment remove* Insert a non-interacting segment S_i into an existing interaction I, or remove a segment from the interaction. Corresponds to adding or removing a β-strand from a β-sheet, allowing buildup of β-sheets with more than two β-strands.

The modified algorithm proceeds as follows:

Algorithm 8.4 (MCMC Segmentation).

1) Initialize $(\mathcal{S}^{(0)}, \mathcal{I}^{(0)})$.

2) Iterate:

(a) Draw $k \sim Uniform[1, m^{(t)}]$, *and set*

$$(\mathcal{S}', \mathcal{I}') = \begin{cases} segment\, split(k) & with\ probability\ .5 \\ segment\, merge(k) & otherwise \end{cases}$$

and then set

$$(\mathcal{S}^{(t+1)}, \mathcal{I}^{(t+1)}) = \begin{cases} (\mathcal{S}', \mathcal{I}') & with\ probability\ \rho_{split(k)} \\ (\mathcal{S}^{(t)}, \mathcal{I}^{(t)}) & otherwise \end{cases}$$

(b) With probability .5,

i. draw $j \neq k \sim Uniform[1, m^{(t)}]$

ii. draw $a_{j,k} \sim Uniform[1, |\{alignments_{j,k}\}|]$

iii. set $I^{new} = (\{j, k\}, a_{j,k})$

iv. set

$$(\mathcal{S}^{(t+1)}, \mathcal{I}^{(t+1)}) = \begin{cases} (\mathcal{S}^{(t)}, \mathcal{I}^{(t)} \cup \{I^{new}\}) & with\ prob.\ \rho_{join(j,k)} \\ (\mathcal{S}^{(t)}, \mathcal{I}^{(t)}) & otherwise \end{cases}$$

Otherwise,

i. draw $I = (j, k, a_{j,k}) \sim Uniform[1, |\mathcal{I}^{(t)}|]$

ii. set

$$(\mathcal{S}^{(t+1)}, \mathcal{I}^{(t+1)}) = \begin{cases} (\mathcal{S}^{(t)}, \mathcal{I}^{(t)} \backslash I) & \text{with prob. } \rho_{join(j,k)} \\ (\mathcal{S}^{(t)}, \mathcal{I}^{(t)}) & \text{otherwise} \end{cases}$$

(c) For $k = 1 : m^{(t+1)}$, *draw* $S_k^{(t+1)} \sim P(S_k \mid m^{(t+1)}, S_{j<k}^{(t+1)}, S_{j>k}^{(t)}, T_{j<k}^{(t+1)}, T_{j\geq k}^{(t)})$

(d) For $k = 1 : m^{(t+1)}$, *draw* $T_k^{(t+1)} \sim P(T_k \mid m^{(t+1)}, S_{j\leq k}^{(t+1)}, S_{j>k}^{(t)}, T_{j<k}^{(t+1)}, T_{j>k}^{(t)})$

Here $\{alignments_{j,k}\}$ represents the set of possible interaction parameters between the j^{th} and k^{th} segments. (E.g. the set of strand-register alignments between two β-strands.) Of course other more complicated transitions may be imagined, but the general flow of the algorithm remains similar these pose no problem so long as individual transitions are constructed in a reversible manner.

Initialization of Algorithm 8.4 may be done at random or using the results of precomputation. A reasonable approach, taken in Chapter 9 is to initialize $\mathcal{I} = \emptyset$ and $\mathcal{S}$ to the $\mathcal{S}_{MAP}$ obtained under the simpler independent segment model (6.1) as described in Chapter 4.

The following result formalizes the statement that this algorithm provides a basis for posterior inference in SSIMs:

Lemma 8.2. *The Markov chain constructed by Algorithm 8.4 has invariant distribution given by (6.2).*

Proof: Again, this follows by construction. ◇

Hence Algorithm 8.4 provides a Monte Carlo approximation scheme for computing functionals of the posterior distribution $P(\mathcal{S}, \mathcal{I} \mid R)$ under the class of SSIM models (6.2). In particular, predictive quantities of interest such as $Struct_{MAP}$ and $Struct_{Mode}$ can be computed approximately via this algorithm.

I have therefore established a computational machinery for dealing with the generalized class of SSIM models, analogous to the exact algorithms of Chapter 4 for inference with SSM models. Having provided the tools for inference and prediction

with this generalized class of segment-interaction models, the next and final chapter of this dissertation will evaluate the SSM and SSIM models developed in earlier chapters by applying them to prediction of protein structure data.

Chapter 9

Evaluation

This chapter presents an evaluation of the methodology developed in this dissertation, by applying it to prediction of protein structure data obtained from databases of experimentally determined protein structures. I describe the evaluation of both the SSM models and algorithms developed in Chapters 4, 6, and 7, and the SSIM models and algorithms developed in Chapters 5, 6, and 8.

I begin by discussing well-established standards for evaluation of secondary structure prediction, and present the results of experiments evaluating the SSM methodology of Chapter 4 according to these criteria. I show that the Bayesian segmentation algorithm using the SSM models performs at the level of the best-published results in the field.

I then present examples and experiments demonstrating the SSIM methodology developed in Chapter 5. I provide evidence that the $Struct_{MAP}$ and $C_{i,j}$ matrix predictors defined in Section 5.3 can yield informative estimates of β-sheet topology and β-strand contact maps respectively, by application to example protein sequences. I briefly discuss the difficulty of systematically evaluating contact map predictions across large numbers of proteins.

9.1 Issues in evaluating predictions of protein structure

Methodology for accurate evaluation of secondary structure predictions has progressed substantially since early work showing that many traditional methods had far overestimated their predictive accuracies (Nishikawa, 1983). A number of standards have emerged for measuring predictive accuracy in a relatively unbiased fashion. I describe the relevant issues here.

9.1.1 Data sets

Early experiments were often performed on small datasets using homologous structures, due to the lack of available data. Often prediction error rates were computed on training data, leading to upwardly biased estimates of accuracy. It is now standard in the field to report leave-one-out or 10-fold cross-validation results on a carefully screened database of experimental structures (Rost and Sander, 1993b; Frishman and Argos, 1996; Salamov and Solovyev, 1997; Schmidler et al., 2000) and algorithms not subjected to such evaluation are no longer publishable. The database used for evaluation is typically restricted to high-resolution (< 3Å) X-ray crystallographic structures, without identifiable sequence homology ($< 25\%$ sequence identity). Construction of appropriate datasets is greatly facilitated by the availability of graph algorithms which find maximal such datasets (Heringa et al., 1992; Hobohm and Sander, 1994) over the entire Brookhaven Protein Data Bank (PDB) (Bernstein et al., 1977). Two such datasets are used for the results reported in this dissertation:

- *Dataset #1* - The OBSTRUCT program (Heringa et al., 1992) was used in November 1997 to create a maximal non-redundant ($< 25\%$ sequence identity) set of high-resolution (< 2.5Å) globular protein structures from the PDB. From this set, structures classified as membrane proteins by SCOP (Murzin et al., 1995) were removed, as were sequences less than 50 amino acids in length, leaving 451 proteins.

- *Dataset #2* - The PDB-SELECT algorithm (Hobohm and Sander, 1994) was

used in May 1998 to obtain a maximal non-redundant ($< 25\%$) set of high-resolution (< 3Å) structures from the PDB. These 685 proteins were reduced to 660 by removal of membrane proteins and and those for which DSSP produced no output.

The PDB continues to grow at a rapid rate, making larger datasets continually available. This growing data resource allows more accurate estimation of model parameters, as well as fitting of more complicated models. The intent of the experiments shown here is to demonstrate the basic approach of Bayesian segmentation and segment-interaction models for protein structure prediction, and not necessarily to obtain optimal performance. Approaches to maximizing predictive performance given available data by applying automated model selection procedures are suggested in Section 10.1.

9.1.2 Gold standard definition of secondary structure

Given the 3D atomic coordinates of a protein structure as determined by X-ray crystallography, a gold-standard definition of secondary structure is needed. Several algorithms exist for automatic secondary structure assignment from coordinates, using ϕ/ψ angles, putative hydrogen bonding patterns, or combinations of the two.

Significant variation exists among the assignments resulting from different algorithms (Colloc'h et al., 1993); common disagreements include exact boundaries for α-helices and β-strands, and assignments for very short secondary structure segments. Here I use the DSSP algorithm (Kabsch and Sander, 1983) for gold standard assignments, by far the most common choice in the secondary structure prediction literature. DSSP assignments are used for automated annotation of PDB structures. These assignments are adjusted as suggested by the literature (Frishman and Argos, 1996) to restrict the minimum β-strand length to 3, and the minimum α-helix length to 5.

For the segment-based models developed here, DSSP assignments raise a difficulty in the treatment of helix end positions. DSSP helix assignments exclude first and last hydrogen bonded residues, so the N_{cap} and C_{cap} positions are typically omitted.

Effects of this can be observed in Figure 2.5. As described in Chapter 2, these N- and C- terminal positions provide important signals for identifying α-helices in protein sequences. To partially account for this, I allow the segment transition term in (4.1) to depend on the last residue of the previous segment.

9.1.3 Accuracy measures

Given a non-redundant, high-resolution database and a gold standard secondary structure assignment, prediction accuracies may be estimated by cross-validation experiments. It is common to report several quantities from these experiments:

- *Overall 3-state accuracy (Q_3)*:
  The most commonly reported measure of secondary structure prediction accuracy is the percentage of individual amino acids in the database assigned to the correct state. Accuracies quoted in Chapter 3 for existing algorithms are Q_3 values. An undesirable property of Q_3 is the dependence on underlying database composition; however it remains attractive as a single numerical summary of overall accuracy.

- *Sensitivity (Q_α^{obs}, Q_β^{obs}, Q_L^{obs})*:
  The sensitivity of predicting α-helical positions (Q_α^{obs}) is estimated by the the percentage of helical residues predicted to be helical. Sensitivity is calculated as $\frac{TP}{TP+FN}$ where TP = true positives and FN = false negatives.

- *Positive predictive value (Q_α^{pred}, Q_β^{pred}, Q_L^{pred})*:
  The positive predictive value (PPV) for α-helices (Q_α^{pred}) is estimated by the percentage of predicted helical residues which are truly helical. PPV is calculated as $\frac{TP}{TP+FP}$, where FP = false positives.

- *Matthew's correlation (C_α, C_β, C_L)*:
  The correlation coefficient introduced by (Matthews, 1975) is insensitive to the

underlying database composition, and is defined by

$$C_i = \frac{(TP)(TN) - (FN)(FP)}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \quad \text{for} \quad i \in \{\alpha, \beta, L\} \tag{9.1}$$

where TN = true negatives.

Several other accuracy measures have been proposed for secondary structure prediction, including attempts to measure prediction of secondary structure *segments* rather than individual residues (Taylor, 1984; Presnell et al., 1992; Rost et al., 1994). However, such a measure is difficult to define satisfactorily, and none have achieved common usage in the literature. Here I provide results only for those quantities given above, which serve as a standard basis for comparison to other secondary structure prediction algorithms.

9.2 Evaluation of stochastic segment models

The previous section discussed widely accepted experimental methodology for evaluation of protein secondary structure prediction algorithms. In this section I describe the evaluation of the secondary structure prediction approaches developed in this dissertation according to these criteria.

Table 9.1 shows the results of applying the SSM models of Chapter 4 to *Dataset*#1, using the algorithms of Chapter 7. Under the semi-Markov segmentation prior, the marginal mode predictor is seen to significantly outperform the MAP segmentation by the Q_3 measure, as expected from the optimality arguments given in Chapter 6. The marginal mode predictor achieves an accuracy of 68.8%, competitive with the best published results as described in Section 3.1.6.

By way of example, Figure 9.1 shows a typical sequence prediction, where we see that segment endpoints are the regions of highest uncertainty, as expected. These positions also reflect the highest variability among different structure assignment algorithms (Colloc'h et al., 1993).

	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
Mode	68.8	64.0 (69.7)	46.0 (61.0)	81.0 (70.5)	.54	.43	.47
MAP	64.2	67.3 (61.8)	23.3 (61.3)	79.1 (65.9)	.49	.30	.38

Table 9.1: *Dataset*#1 cross-validation results for Bayesian segmentation algorithm using SSM models. Results are given for MAP segmentation and marginal mode predictors using a semi-Markov prior. Data from (Schmidler et al., 2000)
.

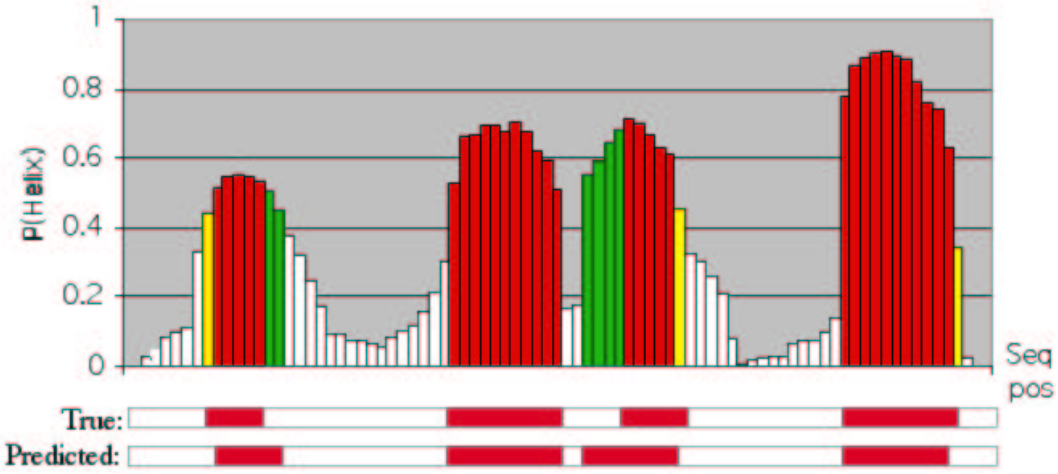


Figure 9.1: Prediction of secondary structure for Cytochrome C. Bars indicate predicted probability of α-helical structure. Positions in red (white) are correctly predicted to be in a helical (coil) conformation. Positions in green are over-predicted (true structure coil, predicted structure helix). Positions in yellow are under-predicted (true structure helix, predicted structure coil). Over- and under- predictions occur primarily near segment endpoints, and are predicted at lower probability in general.

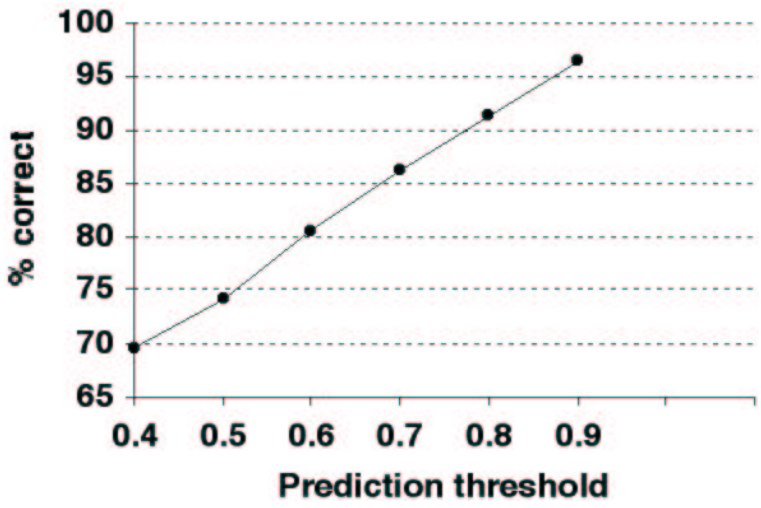


Figure 9.2: Plot of predictive accuracy versus probability assigned to prediction (*Dataset*#1). The strong correlation indicates accurate estimation of prediction uncertainty at each sequence position.

In addition to accuracy, the thesis of this dissertation requires that the methodology provide accurate estimates of prediction uncertainty. The success of the Bayesian framework developed here according to this criteria is clearly shown in Figure 9.2, which plots the empirical accuracy for residues predicted at a series of probability thresholds. As can be seen from the strong correlation, a clear advantage of the explicit probabilistic approach developed in this dissertation is accuracy of estimated prediction confidence at each position. At a threshold prediction probability of 0.6, predictions are made for 58% of positions and achieve an accuracy of 80.6%. At a threshold probability of 0.8, the algorithm achieves an accuracy of 91.4%, but predicts only 21% of positions with this level of confidence. According to (Rost and Schneider, 1998), these threshold percentages indicate that the Bayesian segmentation approach using independent segments performs 6 times as well as other single sequence methods which provide reliability estimates, and methods based on multiple sequence alignments such as PhD (Rost and Sander, 1994) perform only $\frac{7}{6}$ times better than the Bayesian approach applied to single sequences.

Table 9.2 shows the results of experiments using *Dataset*#2. Here a comparison between two alternative segmentation priors is made: the semi-Markov prior (4.2)

Alg & Prior	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
Mode_{SM}	67.9	64.8 (68.0)	44.8 (58.7)	79.1 (70.4)	.53	.41	.46
MAP_{SM}	63.9	70.5 (60.1)	23.1 (60.6)	76.6 (66.7)	.49	.29	.38
Mode_U	50.3	58.1 (53.7)	71.9 (33.2)	37.0 (77.4)	.36	.28	.31
MAP_U	50.2	62.2 (48.1)	59.0 (34.8)	39.5 (72.6)	.32	.26	.28

Table 9.2: *Dataset*#2 cross-validation results for Bayesian segmentation algorithm using SSM models. Results are given for MAP segmentation and marginal mode predictors for two priors: semi-Markov (SM) and uniform (U) as defined in Chapter 6.

used previously, and a uniform prior (6.22) yielding a maximum likelihood segmentation. The uniform segmentation prior is seen to yield significantly lower predictive performance, demonstrating the important role played by the prior in the Bayesian segmentation algorithm. This may be explained in part by the 1-to-1 growth of the number of "parameters to be estimated" (i.e. the number of secondary structure assignments) with the number of "data points" (i.e. sequence length). Thus prior specification is an important part of the modelig process, and further experimentation with alternative priors may yield further improvements. Figure 9.3 shows that prediction confidence once again correlates strongly with prediction accuracy.

9.3 Evaluation of stochastic segment interaction models

The previous sections were concerned with evaluation of SSM models as predictors of protein secondary structure. In this section I consider the use of SSIM models for prediction of both secondary structure and β-sheet contact maps, as described in Chapters 5 and 6.

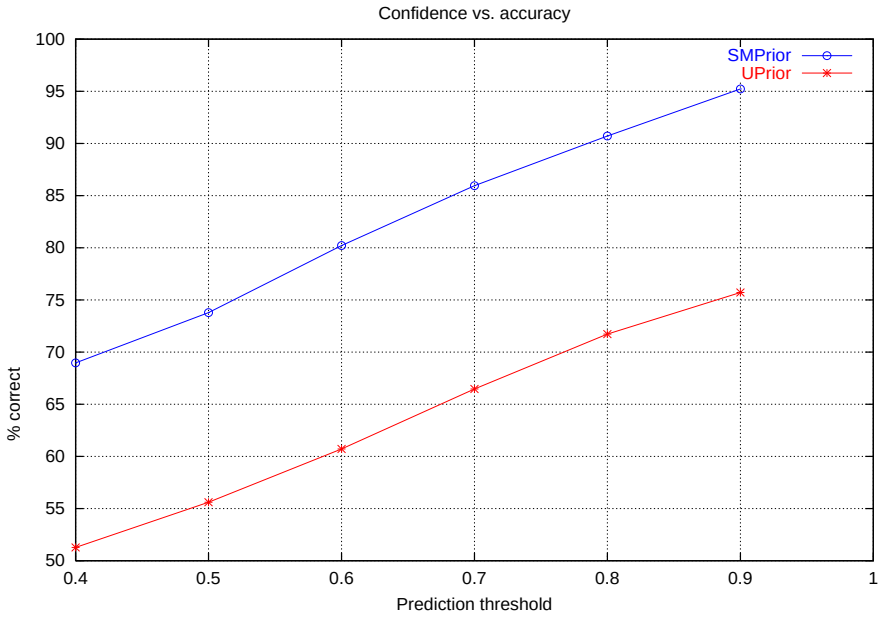


Figure 9.3: Plot of predictive accuracy versus probability assigned to prediction (*Dataset*#2), for semi-Markov and Uniform priors. The strong correlation represents accurate estimation of prediction uncertainty at each position in a sequence.

9.3.1 Impact of segment interaction models on secondary structure prediction

I begin by evaluating the impact of including non-local segment interactions into the Bayesian framework on the resulting secondary structure prediction accuracy. This can be done by repeating the cross-validation experiments described in the previous section using, the MAP and Mode predictors of segmentation defined in Chapters 5 and 6 as predictors of protein secondary structure. This will measure the effect of adding non-local β-sheet models to the accuracy of predicting secondary structure.

As described in Chapters 5 and 6, the chosen prior on segment interactions may have a dramatic impact on the marginal distribution over segmentations. In particular, uniform priors such as (5.11) highly bias the marginal distribution towards β-strand, and the resulting predictors are not useful for secondary structure prediction. Hence I consider only priors which preserve the marginal distribution (5.15). The experiments described in this section utilize a segment-interaction prior of the

	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
SSM	68.0	65.1 (68.1.)	44.9 (58.7)	79.2 (70.6)	.53	.41	.46
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 9.3: Comparison of secondary structure prediction results for independent segment models (SSM) vs joint-segment models for β-sheets (SSIM). Also shown are results under the SSM model using MCMC inference (SSM_{MC}) in place of exact algorithms. Data consists of 100 randomly chosen proteins from *Dataset*#2.

form (5.23) as described in Chapter 5.

Table 9.3 shows the results of secondary structure prediction under the SSIM model defined by (5.9) and (5.23) on 100 randomly-sampled proteins from *Dataset*#2, compared to the predictions for the same proteins obtained from the SSM models defined by (4.3), (4.4), and (4.5) used in the previous sections. The SSIM model predictions were made using the MCMC algorithm developed in Chapter 8; each simulation was run for 25,000 iterations and the first 1,000 "burn-in" samples discarded. It can be seen that the result is a small decrease in performance, resulting from a higher sensitivity (but lower specificity) in predicting β-strand positions. 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.

9.3.2 Evaluation of tertiary contact prediction

SSIM models are of interest not only for the possibility of improving marginal segmentation accuracy (e.g. secondary structure prediction), but also for the potential to predict interactions (e.g. β-sheet contact maps) as described in Chapter 6. In fact, for application to protein structure prediction, the latter is of significantly greater interest. Predicting non-local contacts in protein sequences would constitute a major step beyond traditional secondary structure prediction. In this Section I report on experiments aimed at evaluating the ability of the non-local β-sheet models developed in Chapter 5 to predict non-local β-sheet contacts and topology in real proteins.

Examples

The SSIM model for β-sheet interactions (5.9,5.23) used in the previous section may easily be applied to predict β-sheet contacts using the approach developed in Section 5.3. In this section I demonstrate this application to a range of example proteins.

Figure 9.4 shows the true and predicted β-strand contact maps for two example protein sequences, bovine pancreatic trypsin inhibitor (BPTI), and flavodoxin. BPTI is a small protein with a two-strand anti-parallel β-sheet (a β-hairpin); flavodoxin is larger with an interior 5-strand parallel β-sheet. Contact maps are calculated using the marginal mode estimator (5.31) to compute probability of contact for all pairs of sequence positions. Figure 9.5 shows the results obtained under a model restricted to consider strand pairings of only the correct orientation (parallel for BPTI, anti-parallel for flavodoxin). It can be seen that the latter exhibits less uncertainty. In both cases the strand regions of the sequence are identified with high probability, and true contacts assigned relatively high probability. However various non-native contacts are also assigned comparable probabilities, making predictions largely non-specific. Proper alignment of register between paired strands is a noticeable source of noise. Flavodoxin shows significantly more uncertainty in the predicted contacts, possibly because it contains more β-strands for possible pairings, and these strands are buried in a hydrophobic core, potentially making specific side chain interactions less dominant.

Figure 9.6 shows the results of applying the model to predict β-sheet contacts on the sample of 100 proteins from *Dataset*#2 analyzed in Table 9.3. The ROC curve plots sensitivity vs. (1-specificity) for a range of thresholds for the probability of a predicted contact. It can be seen that it is possible to obtain approximately %80 sensitivity and %75 specificity simultaneously, or %90 specificity with a sensitivity of about %63. %90 sensitivity occurs at only about %56 specificity. In order to explore the implications of these results, I consider in the next section a set of "case studies".

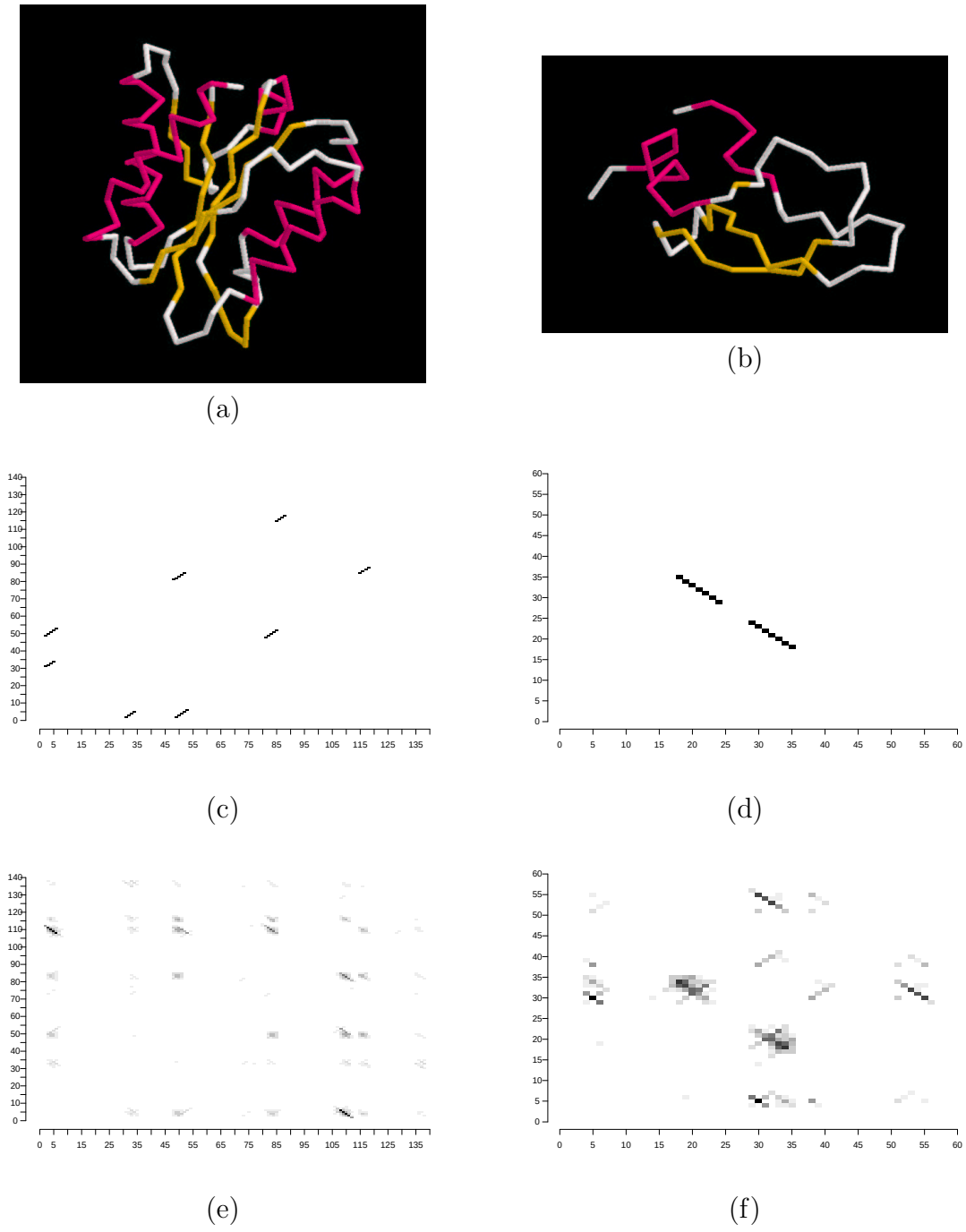


Figure 9.4: β-sheet contact map prediction for BPTI (5pti) and flavodoxin (5nul). Shown are (a,b) X-ray crystallographic structure obtained from Protein Data Bank, (c,d) true contact map derived from crystal structure, and (e,f) probability of contacts predicted from sequence. Contact map axes represent position in sequence. Shading of pixels (x,y) indicates predicted probability of residues x,y forming contacts within a β-sheet.

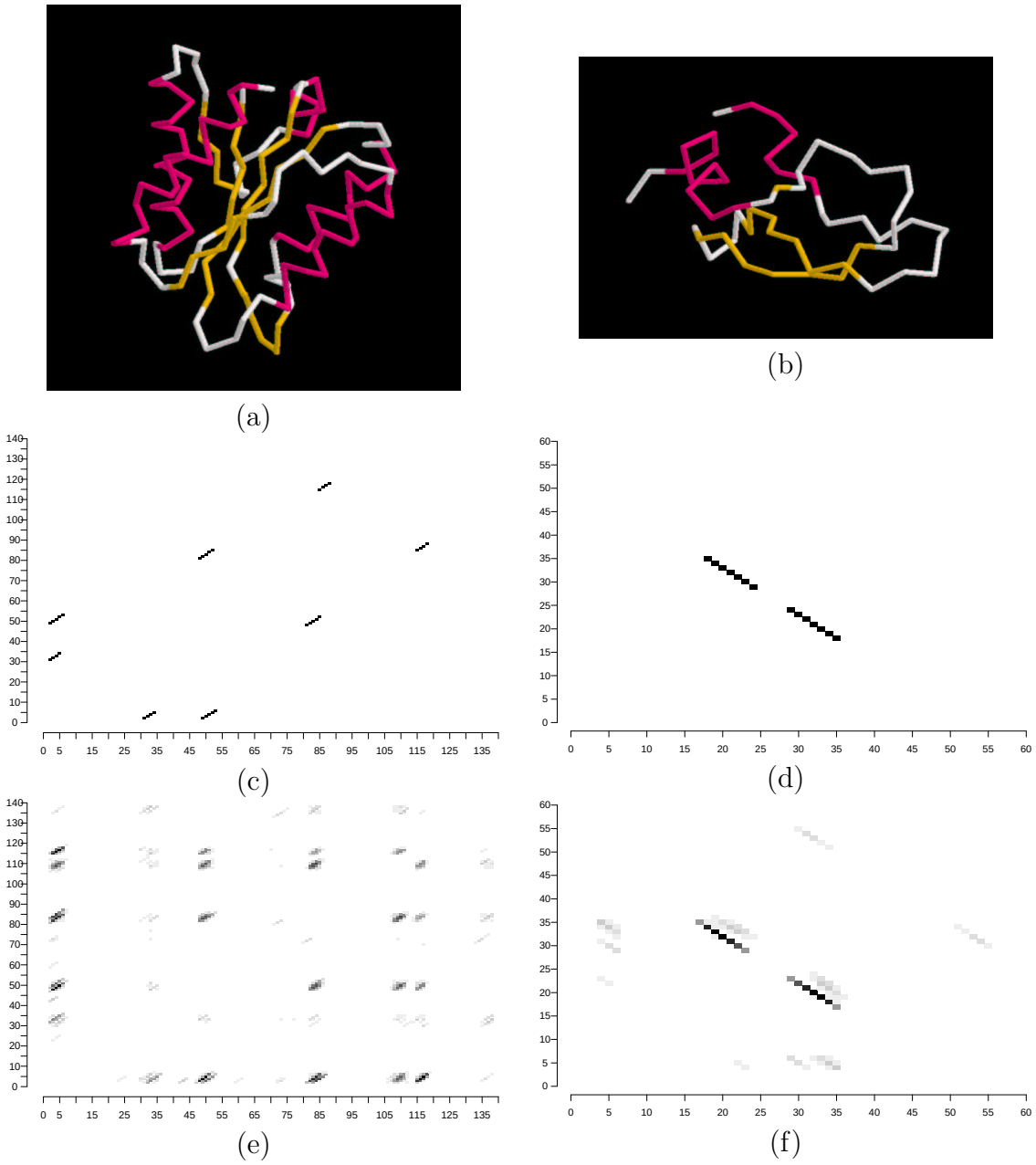


Figure 9.5: β-sheet contact map prediction for BPTI (5pti) and flavodoxin (5nul), under model restricted to correct strand orientation. Shown are (a,b) X-ray crystallographic structure obtained from Protein Data Bank, (c,d) true contact map derived from crystal structure, and (e,f) probability of contacts 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.

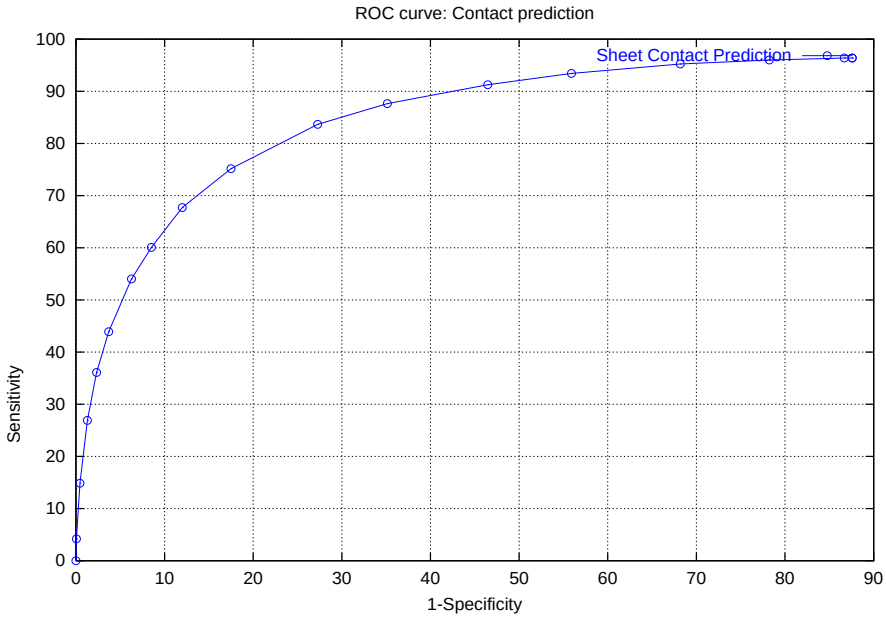


Figure 9.6: ROC curve for β-sheet contact predictions as a function of (log) probability of predicted contact. ROC curve plots Sensitivity vs. (1-Specificity). Data consists of the 100 randomly sampled proteins from *Dataset*#2 analyzed in Table 9.3.

Case studies

In this section I explore specific examples the application of segment-interaction models to prediction of β-sheet contacts. The proteins studied here are listed in Table 9.4, and were chosen with the goal of obtaining representatives from each of the major β-containing structural classes defined by SCOP (Murzin et al., 1995).

All β proteins Figures 9.7 and show predictions for three proteins in the *All* β class: an immunoglobulin (1igt_A), a heat-shock protein (1shs_A), and a protease inhibitor (1ecz_A). In each case, many native contacts are predicted with significant probability. However in each case many non-native contacts receive comparable probability, and the overall result is low specificity in the predictions. Significant uncertainty occurs in both strand pairing and pair orientation, while individual strands tend to be located fairly well.

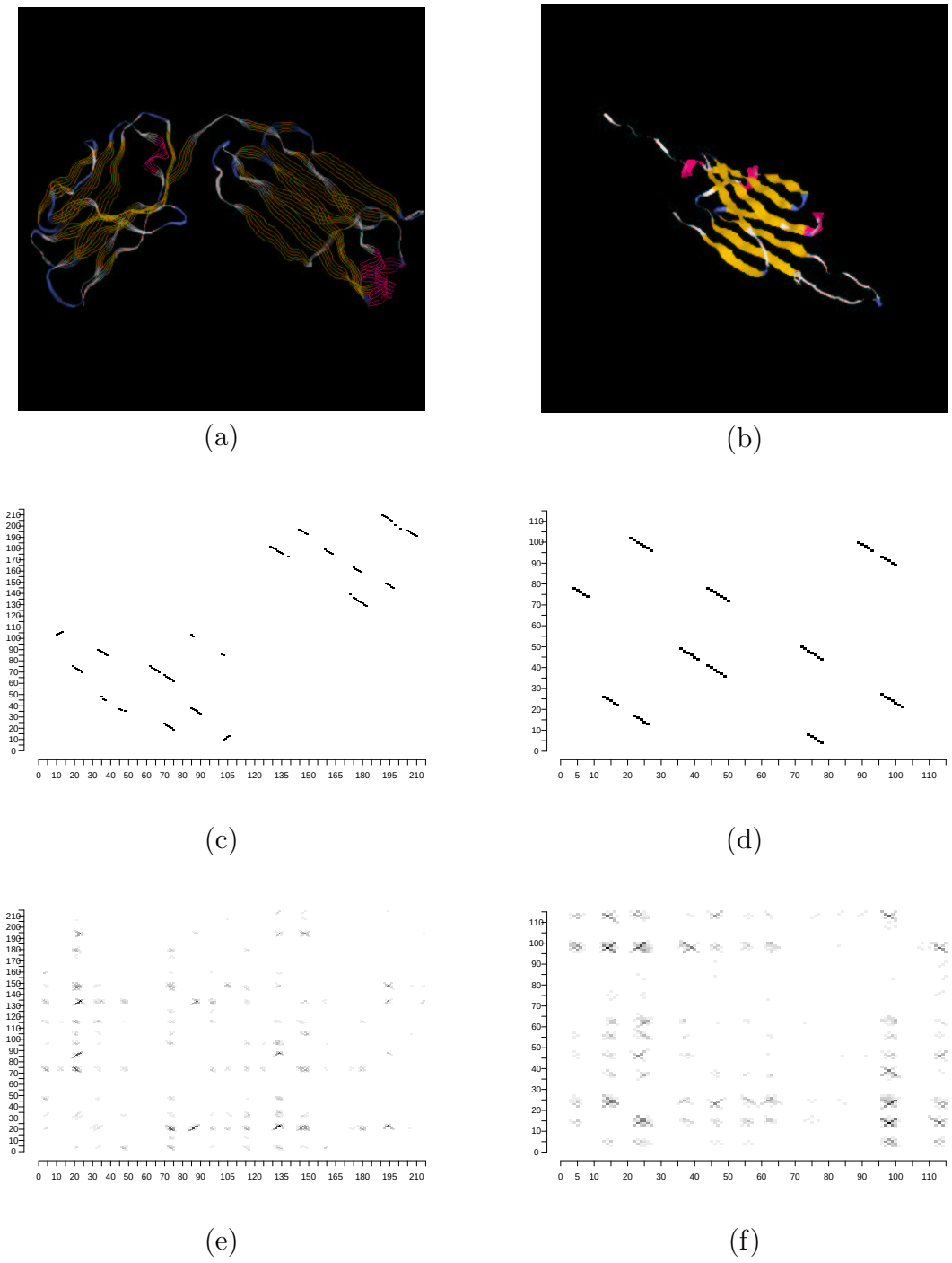


Figure 9.7: β-sheet contact map prediction for two all-β proteins. Shown are (a,c,e) an immunoglobulin (1igt_A) and (b,d,f) a heat-shock protein (1shs_A). Shown are (a,b) X-ray crystallographic structure obtained from Protein Data Bank, (c,d) true contact map derived from crystal structure, and (e,f) 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.

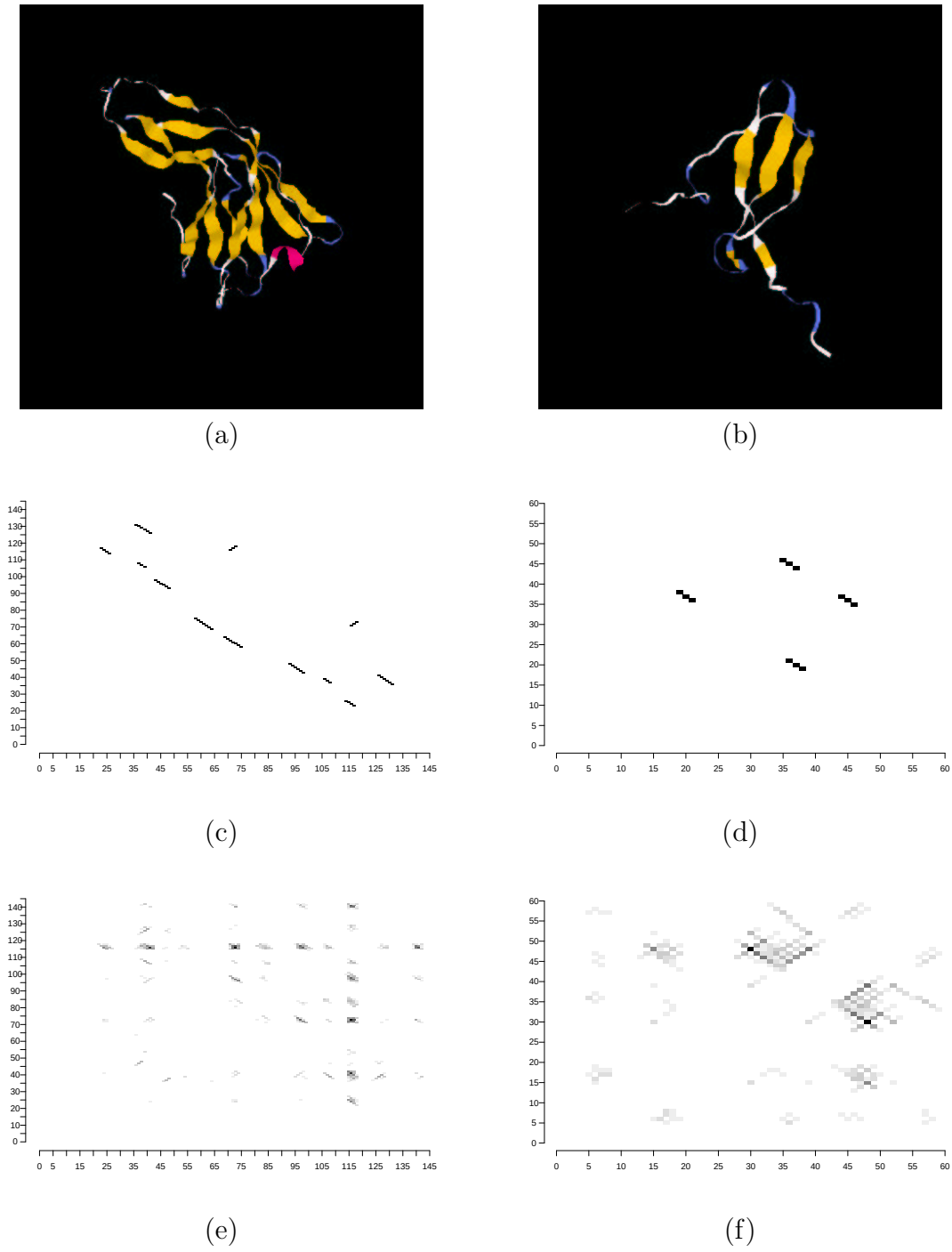


Figure 9.8: β-sheet contact map prediction for one all-β and one small protein. Shown are (a,c,e) a protease inhibitor (1ecz_A) and (b,d,f) heparin-binding growth factor (1mkn_A). Shown are (a,b) X-ray crystallographic structure obtained from Protein Data Bank, (c,d) true contact map derived from crystal 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.

ID	Class	Protein
5pti	Small	BPTI
7pti	Small	BPTI mutant
1igt_A	All β	Immunoglobulin
1shs_A	All β	Heat shock protein
1ecz_A	All β	Protease inhibitor
5nul	α/β	Flavodoxin
1tph_1	α/β	TIM barrel
1rbx	$\alpha+\beta$	RNase A
1b10_A	$\alpha+\beta$	Prion
1mkn_A	Small	Heparin binding growth factor
2bb8	Small	Integrase DNA recombination domain
1aho	Small	Scorpion neurotoxin

Table 9.4: Case study proteins: PDB identifier, SCOP classification, common name, comments.

Small proteins Figures 9.8, 9.9, and 9.10 show predictions for several proteins from the *Small protein* class: BPTI (5pti) and a BPTI mutant (7pti), a heparin-binding growth factor (1mkn_A), a DNA recombination domain from integrase (2bb8), and a scorpion neurotoxin (1aho). In comparison with the larger proteins, these show significantly fewer false predicted contacts. This may be attributed in part to the fewer number of β-strands per sequence, making strand pairing less variable. Alternatively, small proteins may exhibit stronger side-chain correlations as a mechanism for additional stabilization of the tertiary structure.

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. An interesting case is BPTI, where by comparing the mutant structure (7pti) with the native (5pti) we see that the mutant sequence has significantly fewer false predictions. This may be explained by the mutant sequence which substitutes Cys (51) to Ala (51) and Cys (30) to Ala (30). Hence we see that the predictions for BPTI are misled by identifying the disulfide bond (51-30) as a potential strand pairing. This illustrates the limitations of the current model - because β-strand are permitted to interact but not α-helices, the algorithm predicts a β-strand pairing at

(51-30) in order to account for the Cys-Cys bond. When these Cys are removed in the mutant structure, no significant pairings besides the native one are predicted.

α/β proteins Figure 9.11 shows predictions for two proteins from the α/β class: flavodoxin (5nul) and a TIM barrel (1tph_1). Both contain large numbers of β-strands, and while these strands are fairly well identified, little specificity in strand pairing is observed in the predicted maps.

$\alpha+\beta$ proteins Figure 9.12 shows predictions for two proteins from the $\alpha+\beta$ class: ribonuclease A (1rbx) and a prion protein (1b10_A). The ribonuclease prediction shows similar properties to those above - native contacts predicted non-specifically, with significant uncertainty in both strand pairing and pair orientation. The prion structure predicts β-contacts in a non-native region of the sequence, but fails to identify the native contacts. This is interesting in light of the current hypothesis that prion proteins may find a thermodynamically favorable (lower free energy) state than the native state by refold to form significant non-native β-structure.

Remarks

As noted above, low specificity in predicted β-strand contacts is observed in many of the case study proteins. One potential remedy is the introduction of more detailed β-sheet models as described in Section 5.2.4.

Comparison of Figures 9.4 and 9.5 suggests an additional approach. In particular, we may perform the predictions under two separate models each protein analyzed: one in which all β-strand pairings are restricted to parallel orientations, and another with restriction to anti-parallel orientations. The resulting predictions may provide a clearer picture of which strand pairings are spurious, and which are indeed native contacts, as native pairings will not be intermixed with orientation-reversals and other obscuring interactions. While this approach will rule out simultaneous parallel and anti-parallel pairings within a single β-sheet, such pairings are relatively rare in known proteins, and therefore of limited concern. For a given set of interacting segments, or an entire protein, it is also possible to compute formally the marginal

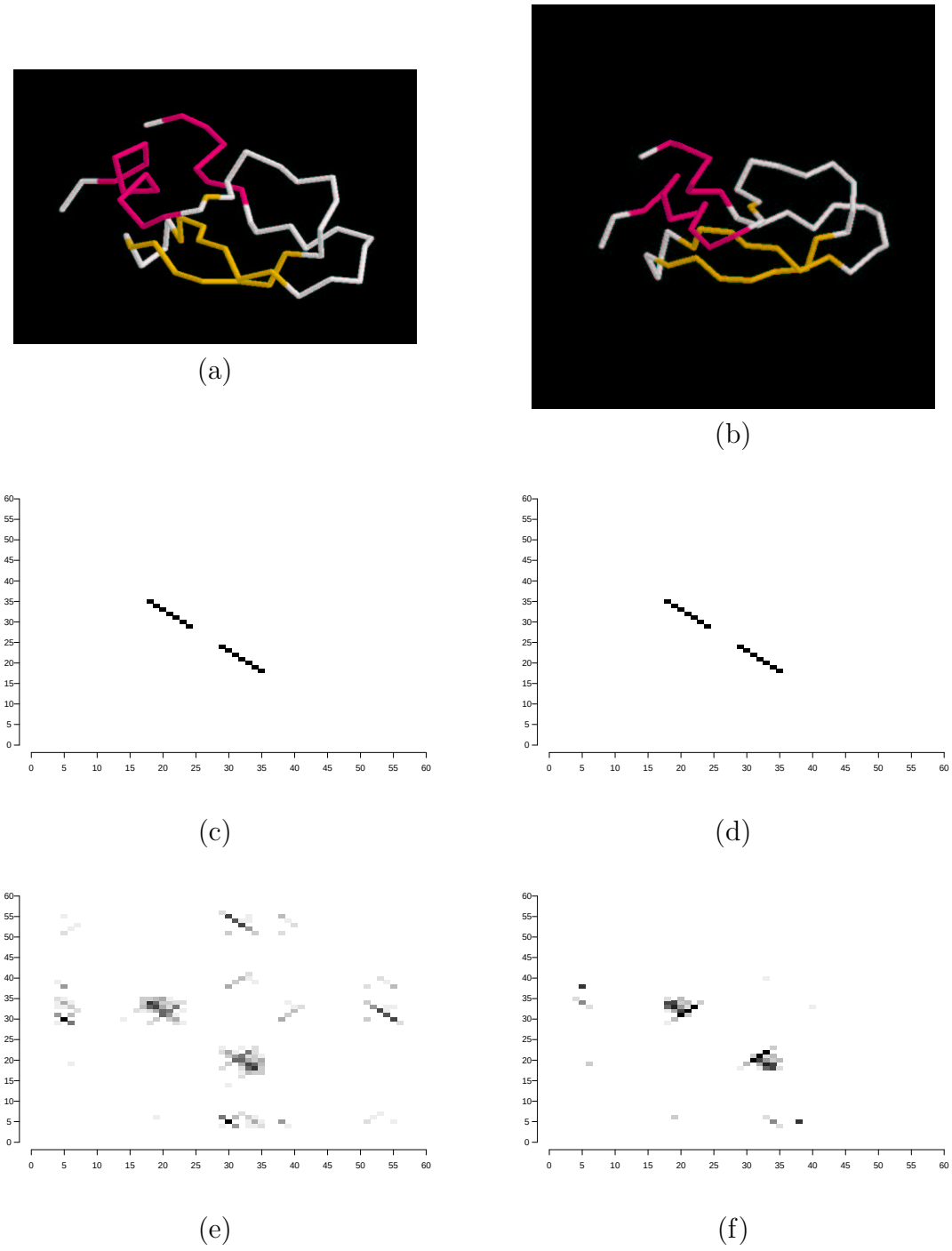


Figure 9.9: β-sheet contact map prediction for two small BPTI proteins. Shown are (a,c,e) BPTI (5pti) and (b,d,f) a BPTI mutant (7pti). Shown are (a,b) X-ray crystallographic structure obtained from Protein Data Bank, (c,d) true contact map derived from crystal structure, and (e,f) probability of contacts 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.

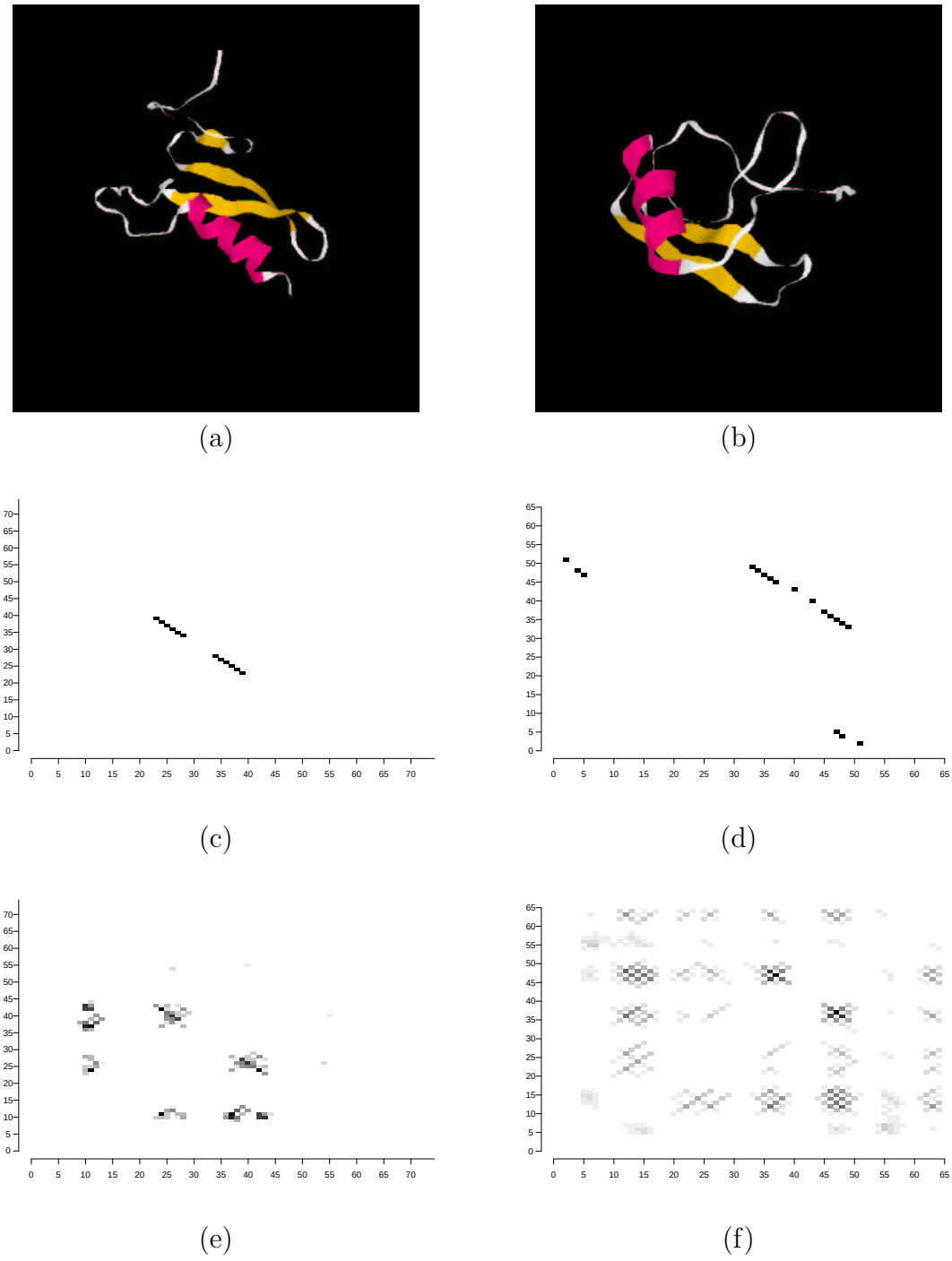


Figure 9.10: β-sheet contact map prediction for two small proteins. Shown are (a,c,e) an integrase DNA recombination domain (2bb8) and (b,d,f) a scorpion neurotoxin (1aho). Shown are (a,b) X-ray crystallographic structure obtained from Protein Data Bank, (c,d) true contact map derived from crystal structure, 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.

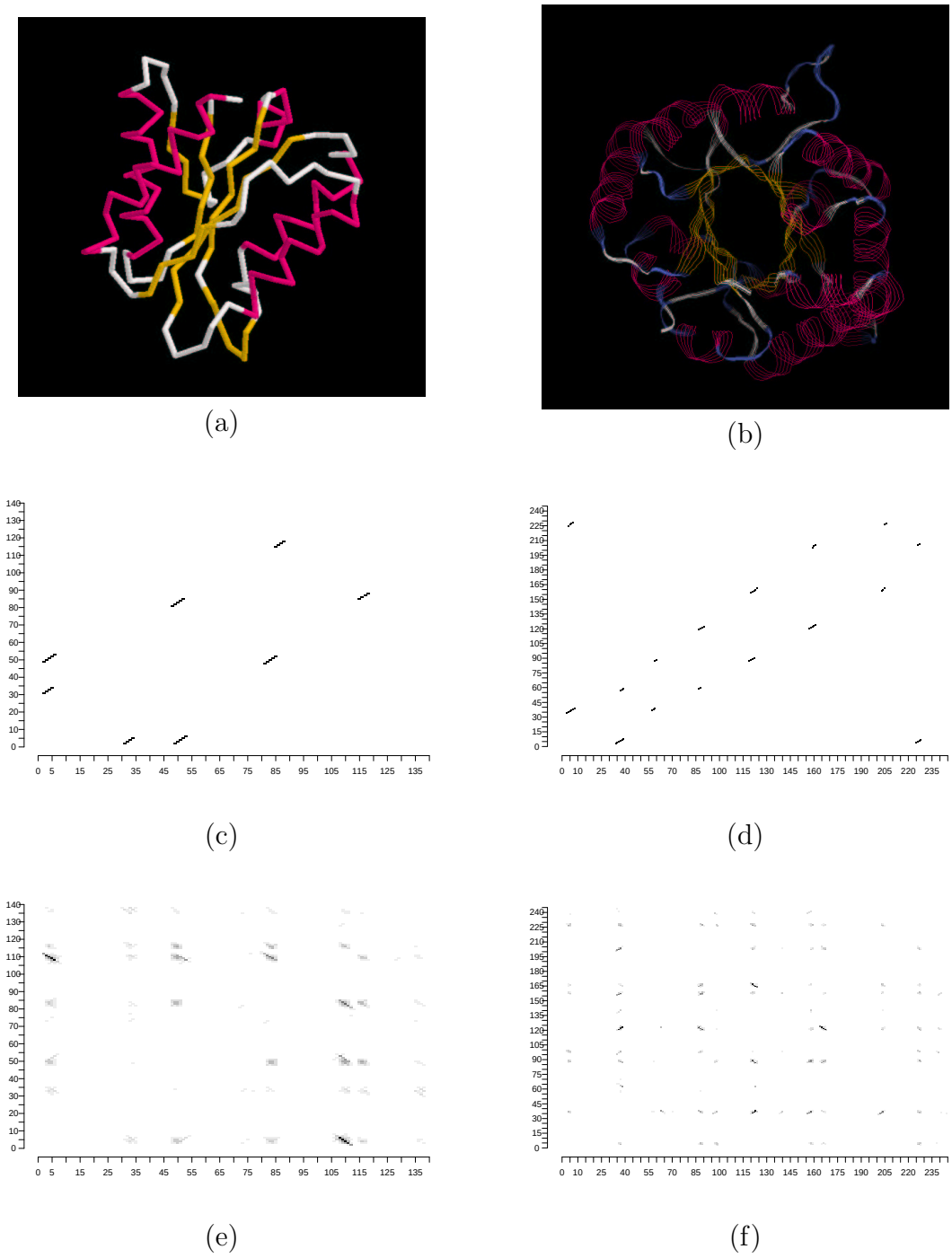


Figure 9.11: β-sheet contact map prediction for two α/β proteins. Shown are (a,c,e) flavodoxin (5nul) and (b,d,f) a TIM barrel (1tph_1). Shown are (a,b) X-ray crystallographic structure obtained from Protein Data Bank, (c,d) true contact map derived from crystal structure, and (e,f) probability of contacts 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.

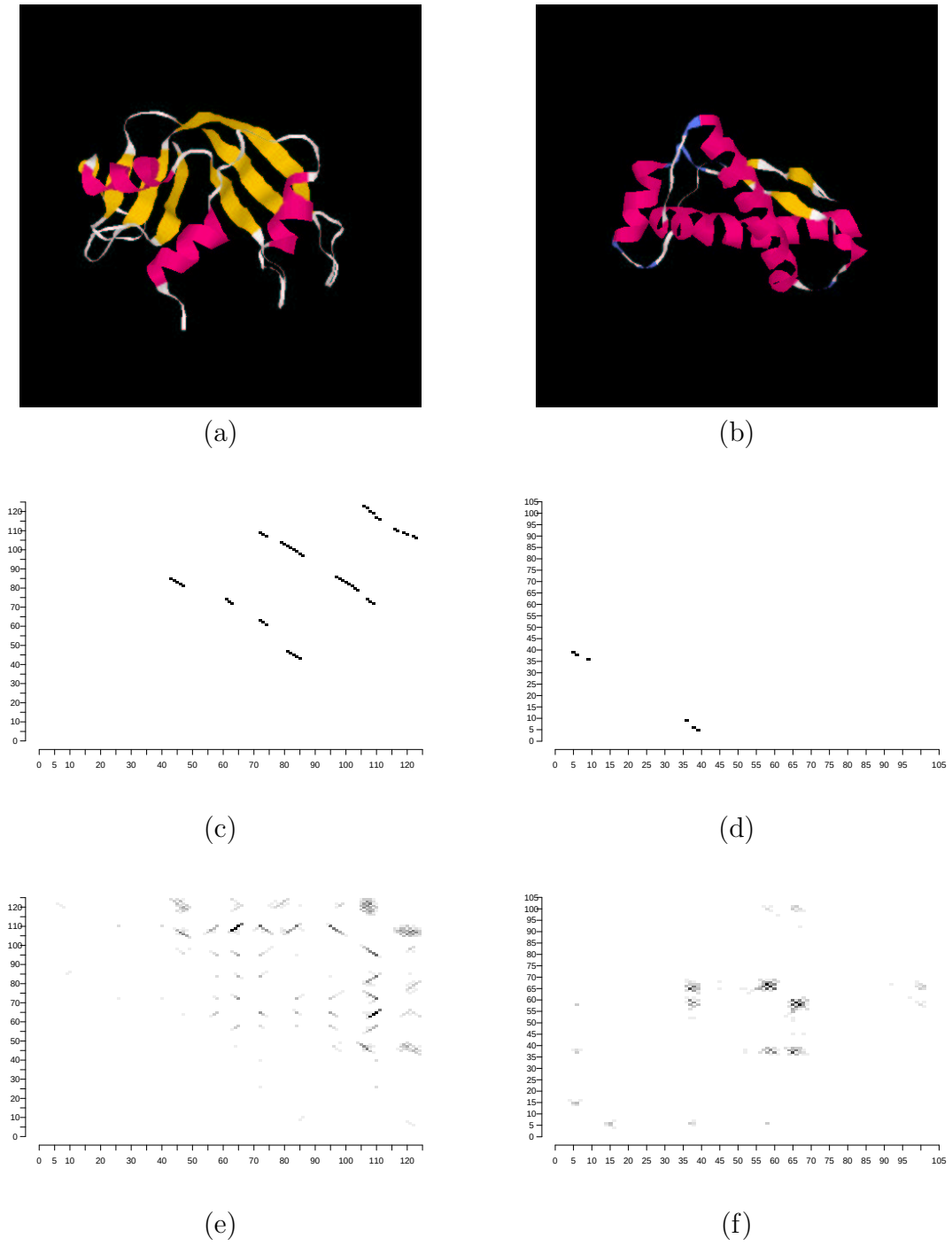


Figure 9.12: β-sheet contact map prediction for two $\alpha+\beta$ proteins. Shown are (a,c,e) a ribonuclease A (1rbx), and (b,d,f) a prion protein (1b10_A). Shown are (a,b) X-ray crystallographic and NMR structures obtained from Protein Data Bank, (c,d) true contact maps derived from experimental structures, 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.

probability under the full model that a particular pairing or set of pairings is parallel vs. anti-parallel. Exploring these options for improving strand-pairing specificity will be the subject of future research.

Chapter 10

Summary and Future Work

In this dissertation proposal, I have developed a novel framework for the prediction of protein structure from amino acid sequence, based on a new class of generalized stochastic models for sequence/structure relationships. I have introduced a Bayesian framework for protein structure prediction, involving a set of joint probability models for sequence-structure mappings based on structural segments. I have described a set of probabilistic models for structural segments characterized by conditional independence of inter-segment positions, developed efficient computational tools for inference and prediction with this class of models, and demonstrated that this approach yields predictive accuracies comparable to the best published methods via extensive cross-validation experiments on experimentally determined structures. In addition, I have shown that such models provide accurate estimates of prediction uncertainty, allowing users to identify regions of each individual protein which may be predicted with especially high accuracy.

I have then gone on to generalize this Bayesian framework to models of non-local interactions in protein sequences, allowing incorporation of factors which contribute significantly to tertiary fold formation. I have shown that computation in this class of models may be performed via Markov chain Monte Carlo algorithms, and have demonstrated this approach by developing a particular set of models for correlated mutations in β-strands which pair to form β-sheets. I have shown via case studies and large cross-validation experiments that this approach can provide predictions of

β-strand contact maps in proteins, providing important information about protein tertiary structure from sequence alone.

10.1 Future work

In this section I briefly describe future directions for the contributions developed in this dissertation. These consist primarily of directions for improvement of the protein structure prediction algorithms developed in this dissertation. More broadly, other statistical applications of the SSM and especially the novel SSIM models developed in this dissertation remain to be pursued. Of particular interest are problems in time-series analysis or other sequential data with long-range dependencies.

10.1.1 Tertiary folding using predicted secondary structure and tertiary β-sheet contacts

The algorithms developed in this dissertation provide probabilistic segmentation of a protein sequence into secondary structure segments. The ability to generate sample segmentations according to their posterior probability may prove to be useful in the continued development of algorithms to predict low-resolution tertiary structures from secondary structure predictions. In particular, by allowing initialization of such algorithms from multiple high-probability regions of conformational space which may be separated by large energy barriers, a much larger class of structural space may be explored.

Moreover, the inclusion of non-local contacts into such algorithms can significantly improve the ability to search relevant regions of conformational space, as described in Section 3.2. Use of the β-sheet contact map predictions developed in Chapter 5 for this purpose may help provide this important information. In particular, if high probability contact predictions can be developed further to achieve high specificity, only very few such high probability contacts may be needed to significantly improve tertiary folding results. Again, the ability to generate predicted β-sheet topologies according to their posterior probability under the full joint model may further aid in

the exploration of conformational space.

In combination with the MCMC methodology discussed in Chapter 8, I hope to use these structural segment predictions to formulate a Bayesian algorithm for full tertiary structure prediction using stochastic dynamics simulations in the space of torsional angles between secondary structure segments. In combination with more explicit physical potentials, these ideas may allow the computation of high-level thermodynamic properties of proteins which are unattainable using standard Monte Carlo or molecular dynamics simulations.

10.1.2 Membrane proteins

An area of particular interest for targeting future work is in application to predicting and simulating the structure of membrane proteins. While the methods developed here have only been applied to globular proteins, membrane proteins make up an important class of targets for protein structure prediction and simulation. Membrane proteins by their nature are difficult to analyze by experimental methods, and are unlikely to yield to large-scale structure determination efforts. In addition, membrane proteins make up almost 50% of pharmaceutically relevant proteins, playing major roles in cellular signaling and transport. It is straightforward to apply the methodology developed in Chapter 4 to the prediction of transmembrane helices in membrane proteins. I hope to explore the use of other structure prediction and Monte Carlo simulation ideas developed in this thesis for application to prediction and simulation of membrane proteins.

10.1.3 Model selection

In Chapter 4 I introduced the general framework of this dissertation and developed a class of segment-based probability models. Example models for α-helices and β-strands were presented in Sections 4.3.1 and 4.3.2. The structure of these models arose from attempts to model known residue dependencies of the type discussed in Chapter 2. However it is also of interest to consider the problem from a purely statistical perspective. In particular, automated model selection techniques may be

applied to determine the form of these segment models which best fits the existing data. Optimization of these models to provide improved predictive performance is an interesting avenue for future research.

Along these lines, several possible improvements to the simplified β-sheet models used in evaluation experiments were suggested in Chapter 5. These may provide directions for improved specificity of the contact predictions obtained in Chapter 9.

10.2 Conclusions

In conclusion, I have developed a Bayesian framework for protein structure prediction and a suite of statistical models and computational tools for implementing this framework. I have evaluated these approaches on large datasets, showing secondary structure prediction accuracies at the level of best-published methods in the field. I have also demonstrated the use of SSIM models for prediction of β-sheet contacts in proteins. These results, while showing low specificity, show promise in developing methods which synthesize multiple sources of information (local and non-local) to predict protein structure from sequence. Finally I have developed a novel class of stochastic models for sequences of random variables with long-range dependency structure, along with a set of computational algorithms for inference with these models, which will be of interest in a broader class of statistical problems beyond protein structure prediction.

Bibliography

Altman, R. B. (1995). A probabilistic approach to determining biological structure: Integrating uncertain data sources. *Int. J. Human-Comp. Studies*, 42:593–616.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181:223–30.

Aqvist, J., Luecke, H., Quioco, F. A., and Warshel, A. (1991). Dipoles localized at helix termini of proteins stabilize charges. *Proceedings of the National Academy of Sciences USA*, 88:2026–2030.

Armstrong, K. M. and Baldwin, R. L. (1993). Charged histidine affects α-helix stability at all positions in the helix by interacting with the backbone charges. *Proceedings of the National Academy of Sciences USA*, 90:11337–11340.

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

Asogawa, M. (1997). β-sheet prediction using inter-strand residue pairs and refinement with Hopfield neural network. In Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C., and Valencia, A., editors, *Proceedings, Fifth International Conference on Intelligent Systems in Molecular Biology*, pages 48–51. AAAI Press.

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

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

Avbelj, F. and Fele, L. (1998). Role of main chain electrostatics, hydrophobic effect and side chain conformational entropy in determining the secondary structure of proteins. *Journal of Molecular Biology*, 279:665–684.

Avbelj, F. and Moult, J. (1995). Role of electrostatic screening in determining protein main chain conformational preferences. *Biochemistry*, 34:755–764.

Bai, Y. and Englander, S. W. (1994). Hydrogen bond strength and β-sheet propensities: The role of a side chain blocking effect. *Proteins: Structure, Function, and Genetics*, 18:262–266.

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

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

Baldwin, R. L. and Rose, G. D. (1999b). Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends in Biochemical Sciences*, 24:77–83.

Barton, G. J. (1995). Protein secondary structure prediction. *Current Opinion in Structural Biology*, 5:372–376.

Benzinger, T. L., Gregory, D. M., Burkoth, T. S., Miller-Auer, H., Lynn, D. G., Botto, R. E., and Meredith, S. C. (1998). Propagating structure of Alzheimer's β-amyloid (10-35) is parallel β-sheet with residues in exact register. *Proceedings of the National Academy of Sciences USA*, 95:13407–13412.

Berger, B. (1995). Algorithms for protein structural motif recognition. *Journal of Computational Biology*, 2:125–138.

Berger, B. and Wilson, D. B. (1995). Improved algorithms for protein motif recognition. In *Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 58–67, San Francisco, CA. ACM-SIAM.

Berger, B., Wilson, D. B., Wolf, E., Tonchev, T., Milla, M., and Kim, P. S. (1995). Predicting coiled coils by use of pairwise residue correlations. *Proceedings of the National Academy of Sciences USA*, 92:8259–8263.

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

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The protein data bank: a computer-based archival file for macro-molecular structures. *Journal of Molecular Biology*, 112:535–542.

Blaber, M., Zhang, X. J., and Matthew, B. W. (1993). Structural basis of amino acid α-helix propensity. *Science*, 260:1637–1640.

Bowie, J. U., Luthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170.

Branden, C. and Tooze, J. (1999). *Introduction to Protein Structure*. Garland, 2^{nd} edition.

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

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. *Nature Genetics*, 23:151–157.

Chandonia, J.-M. and Karplus, M. (1999). New methods for accurate prediction of protein secondary structure. *Proteins: Structure, Function, and Genetics*, 35:293–306.

Chou, P. Y. and Fasman, G. D. (1974a). Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*, 13:211–222.

Chou, P. Y. and Fasman, U. D. (1974b). Prediction of protein conformation. *Biochemistry*, 13:222–245.

Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, 51(1):79–94.

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

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D., and Fletterick, R. J. (1986). Turn prediction in proteins using a pattern-matching approach. *Biochemistry*, 25:266–275.

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., and Mornon, J.-P. (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Engineering*, 6(4):377–382.

Consortium, I. H. G. S. (2001). A physical map of the human genome. *Nature*, 409:934–941.

Cornett, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., and DeLisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *Journal of Molecular Biology*, 195:659–685.

Creamer, T. P. and Rose, G. D. (1992). Side chain entropy opposes α-helix formation but rationalizes experimentally determined helix-forming propensities. *Proceedings of the National Academy of Sciences USA*, 89:5937–5941.

Creamer, T. P. and Rose, G. D. (1994). α-helix-forming propensities in peptides and proteins. *Proteins: Structure, Function, and Genetics*, 19:85–97.

Creamer, T. P. and Rose, G. D. (1995). Interactions between hydrophobic side chains within α-helices. *Protein Science*, 4:1305–1314.

Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties*. W. H. Freeman and Company, 2^{nd} edition.

Dandekar, T. and Argos, P. (1996). Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. *Journal of Molecular Biology*, 256:645–660.

De Alba, E., Rico, M., and Jiménez, M. A. (1997). Cross-strand side chain interactions versus turn conformation in β-hairpins. *Protein Science*, 6:2548–2560.

De Alba, E., Santoro, J., Rico, M., and Jiménez, M. A. (1999). De novo design of a monomeric three-stranded antiparallel β-sheet. *Protein Science*, 8:854–865.

Di Francesco, V., Garnier, J., and Munson, P. J. (1996). Improving protein secondary structure prediction with aligned homologous sequences. *Protein Science*, 5:106–113.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, 29:7133–7155.

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

Doig, A. J. and Baldwin, R. L. (1995). N- and C-capping preferences for all 20 amino acids in α-helical peptides. *Protein Science*, 4:1325–1336.

Doig, A. J., MacArthur, M. W., Stapley, B. J., and Thornton, J. M. (1997). Structures of N-termini of helices in proteins. *Protein Science*, 6:147–155.

Doig, A. J. and Sternberg, M. J. E. (1995). Side chain conformational entropy in protein folding. *Protein Science*, 4:2247–2251.

Doniach, S. and Eastman, P. (1999). Protein dynamics simulations from nanoseconds to microseconds. *Current Opinion in Structural Biology*, 9:157–163.

Duan, Y. and Kollman, P. A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282:740–744.

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. *Nucleic Acids Research*, 22(11):2079–2088.

Eddy, S. R. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, 6:361–365.

Eisenberg, D., Schwarz, E., Komaromy, M., and Wall, R. (1984a). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *Journal of Molecular Biology*, 179:125–142.

Eisenberg, D., Weiss, R. M., and Terwilliger, T. C. (1982). The helical hydrophobic moment: A measure of the amphiphilicity of a helix. *Nature*, 299:371–374.

Eisenberg, D., Weiss, R. M., and Terwilliger, T. C. (1984b). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proceedings of the National Academy of Sciences USA*, 81:140–144.

Eyrich, V. A., Standley, D. M., Felts, A. K., and Friesner, R. A. (1999a). Protein tertiary structure prediction using a branch and bound algorithm. *Proteins: Structure, Function, and Genetics*, 35:41–57.

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

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

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Doughtery, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O., and Venter, J. C. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269:496–512.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison, C. A., and Venter, J. C. (1995). The minimal gene complement of Mycoplasma genitalium. *Science*, 270:397–403.

Frenkel, D. and Smit, B. (1996). *Understanding Molecular Simulation*. Academic Press.

Friesner, R. A. and Gunn, J. R. (1996). Computer simulation of protein folding. *Ann. Rev. Biol. Biomol. Struct.*, 25:315–342.

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

Frishman, D. and Argos, P. (1997). Seventy-five percent accuracy in protein secondary structure prediction. *Proteins: Structure, Function, and Genetics*, 27:329–335.

Garnier, J., Gibrat, J.-F., and Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol*, 266:540–553.

Garnier, J., Osguthorpe, D. J., and Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, 120:97–120.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman & Hall.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

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

Gobel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Genetics*, 18:309–317.

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

Harper, E. T. and Rose, G. D. (1993). Helix stop signals in proteins and peptides: The capping box. *Biochemistry*, 32:7605–7609.

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

Heringa, J., Sommerfeldt, H., Higgins, D., and Argos, P. (1992). OBSTRUCT: A program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. *Computer Applications in the Biosciences*, 8(6):599–600.

Hobohm, U. and Sander, C. (1994). Enlarged representative set of protein structures. *Protein Science*, 3:522–524.

Holley, H. L. and Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences USA*, 86:152–156.

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

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.

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

Hutchinson, E. G. and Thornton, J. M. (1994). A revised set of potentials for β-turn formation in proteins. *Protein Science*, 3:2207–2216.

Huyghues-Despointes, B. M. P., Klingler, T. M., and Baldwin, R. L. (1995). Measuring the strength of side chain hydrogen bonds in peptide helices: The Gln*Asp $(i, i + 4)$ interaction. *Biochemistry*, 34:13267–13271.

Janek, K., Behlke, J., Zipper, J., Fabian, H., Georgalis, Y., Beyermann, M., Bienert, M., and Krause, E. (1999). Water-soluble β-sheet models which self-assemble into fibrillar structures. *Biochemistry*, 38:8246–8252.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, 358:86–89.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637.

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

Kellis, J. T., Nyberg, K., Sali, D., and Fersht, A. R. (1988). Contribution of hydrophobic interactions to protein stability. *Nature*, 333:784–786.

Kim, C. A. and Berg, J. M. (1993). Thermodynamic β-sheet propensities measured using a zinc-finger host peptide. *Nature*, 362:267–270.

Kim, M. K. and Kang, Y. K. (1999). Positional preference of proline in α-helices. *Protein Science*, 8:1492–1499.

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

Klingler, T. M. (1996). *Structural Inference From Correlations in Biological Sequences*. PhD thesis, Stanford University.

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

Kortemme, T., Ramírez-Alvarado, M., and Serrano, L. (1998). Design of a 20-amino acid three-stranded β-sheet protein. *Science*, 218:253–256.

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

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.

Krylov, D., Mikhailenko, I., and Vinson, C. (1994). A thermodynamic scale for leucine zipper stability and dimerization specificity: e and g interhelical interactions. *EMBO Journal*, 13(12):2849–2861.

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.

Leach, A. R. (1996). *Molecular Modelling: Principles and Applications*. Addison Wesley Longman Ltd.

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

Levitt, M. (1978). Conformational preferences of amino acids in globular proteins. *Biochemistry*, 17:4277–4285.

Levitt, M. and Sharon, R. (1988). Accurate simulation of protein dynamics in solution. *Proceedings of the National Academy of Sciences USA*, 85:7557–7561.

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

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

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.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40.

Lyu, P. C., Liff, M. I., Marky, L. A., and Kallenbach, N. R. (1990). Side chain contributions to the stability of α-helical structure in peptides. *Science*, 250:669–673.

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. *Biochimica et Biophysica 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.

McGregor, M. J., Islam, S. A., and Sternberg, M. J. (1987). Analysis of the relationship between side chain conformation and secondary structure in globular proteins. *Journal of Molecular Biology*, 198:295–310.

Mehta, P. K., Heringa, J., and Argos, P. (1995). A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Science*, 4:2517–2525.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092.

Minor, D. L. and Kim, P. S. (1994a). Context is a major determinant of β-sheet propensity. *Nature*, 371:264–267.

Minor, D. L. and Kim, P. S. (1994b). Measurement of the β-sheet-forming propensities of amino acids. *Nature*, 367:660–663.

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. *Proceedings of the National Academy of Sciences USA*, 91:5027–5029.

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

Motwani, R. and Raghavan, P. (1995). *Randomized Algorithms*. Cambridge University Press.

Munson, P. J., Cao, L., Di Francesco, V., and Porrelli, R. (1993a). Semiparametric and kernel density estimation procedures for prediction of protein secondary structure. In *Amer. Statist. Assoc., Stat. Comp. Section*, San Francisco, California.

Munson, P. J., Di Francesco, V., and Porrelli, R. (1993b). Secondary structure prediction using penalized likelihood models. In *25th Symposium on the Interface, Computing Science and Statistics*, volume 25, San Diego, California.

Munson, P. J., Di Francesco, V., and Porrelli, R. (1994). Protein secondary structure prediction using periodic-quadratic-logistic models: Statistical and theoretical issues. In *Twenty-seventh annual Hawaii international conference on system sciences*, pages 375–384. IEEE.

Murzin, A. Z., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540.

Nishikawa, K. (1983). Assessment of secondary structure prediction of proteins: Comparison of computerized Chou-Fasman methods with others. *Biochimica et Biophysica Acta*, 748:285–299.

Ortiz, A. R., Kolinski, A., and Skolnick, J. (1998). Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. *Proceedings of the National Academy of Sciences USA*, 95:1020–1025.

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.

Padmanabhan, S. and Baldwin, R. L. (1994). Tests for helix-stabilizing interactions between various nonpolar side chains in alanine-based peptides. *Protein Science*, 3:1992–1997.

Padmanabhan, S., Marqusee, S., Ridgeway, T., Laue, T. M., and Baldwin, R. L. (1990). Relative helix-forming tendencies of nonpolar amino acids. *Nature*, 344:268–270.

Pauling, L. and Corey, R. B. (1951). The pleated sheet, a new layer configuration of polypeptide chains. *Proceedings of the National Academy of Sciences USA*, 37:251–256.

Pauling, L., Corey, R. B., and Branson, H. R. (1951). The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences USA*, 37:205–211.

Petukhov, M., Munoz, V., Yumoto, N., Yoshikawa, S., and Serrano, L. (1998). Position dependence of non-polar amino acid intrinsic helical propensities. *Journal of Molecular Biology*, 278:279–289.

Presnell, S. R., Cohen, B. I., and Cohen, F. E. (1992). A segment-based approach to protein secondary structure prediction. *Biochemistry*, 31:983–993.

Presta, L. G. and Rose, G. D. (1988). Helix signals in proteins. *Science*, 240:1632–1641.

Prusiner, S. B. (1997). Prion diseases and the BSE crisis. *Science*, 278:245–251.

Prusiner, S. B. (1998). Prions. *Proceedings of the National Academy of Sciences USA*, 95:13363–13383.

Qian, H. (1996). Prediction of α-helices in proteins based on thermodynamic parameters from solution chemistry. *Journal of Molecular Biology*, 256:663–666.

Qian, N. and Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology*, 202:865–884.

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

Reif, F. (1965). *Fundamentals of Statistical and Thermal Physics*. McGraw Hill.

Richardson, J. S. and Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of α-helices. *Science*, 240:1648–1652.

Riis, S. K. and Krogh, A. (1996). Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *Journal of Computational Biology*, 3(1):163–183.

Rost, B. and Sander, C. (1993a). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Sciences USA*, 90:7558–7562.

Rost, B. and Sander, C. (1993b). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232:584–599.

Rost, B. and Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Structure, Function, and Genetics*, 19:55–72.

Rost, B., Sander, C., and Schneider, R. (1994). Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology*, 235:13–26.

Rost, B. and Schneider, R. (1998). Pedestrian guide to analysing sequence databases. In *Core Techniques in Biochemistry*. Springer, Heidelberg.

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.

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

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. *Nucleic Acids Research*, 22(23):5112–5120.

Salamov, A. A. and Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *Journal of Molecular Biology*, 247:11–15.

Salamov, A. A. and Solovyev, V. V. (1997). Protein secondary structure prediction using local alignments. *Journal of Molecular Biology*, 268:31–36.

Sandberg, W. S. and Terwilliger, T. C. (1989). Influence of interior packing and hydrophobicity on the stability of a protein. *Science*, 245:54–57.

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

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

Shalongo, W. and Stellwagen, E. (1995). Incorporation of pairwise interactions into the Lifson-Roig model for helix prediction. *Protein Science*, 4:1161–1166.

Sibanda, B. L., Blundell, T. L., and Thornton, J. M. (1989). Conformation of β-hairpins in protein structures. a systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *Journal of Molecular Biology*, 206(4):759–777.

Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Current Opinion in Structural Biology*, 5:229–235.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 55(1):3–23.

Smith, C. K. and Regan, L. (1995). Guidelines for protein design: The energetics of β-sheet side chain interactions. *Science*, 270:980–982.

Smith, C. K., Withka, J. M., and Regan, L. (1994). A thermodynamic scale for the β-sheet forming tendencies of the amino acids. *Biochemistry*, 33:5510–5517.

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

Solovyev, V. V. and Salamov, A. A. (1994). Predicting α-helix and β-strand segments of globular proteins. *Computer Applications in the Biosciences*, 10(6):661–669.

Stapley, B. J., Rohl, C. A., and Doig, A. J. (1995). Addition of side chain interactions to modified Lifson-Roig helix-coil theory: Application to energetics of phenylalanine-methionine interactions. *Protein Science*, 4:2383–2391.

Sternberg, M. J. E. (1996). *Protein Structure Prediction: A Practical Approach*. IRL Press.

Stolorz, P., Lapedes, A., and Xia, Y. (1992). Predicting protein secondary structure using neural net and statistical methods. *Journal of Molecular Biology*, 225:363–377.

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.

Street, A. G. and Mayo, S. L. (1999). Intrinsic β-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proceedings of the National Academy of Sciences USA*, 96:9074–9076.

Stryer, L. (1995). *Biochemistry*. W. H. Freeman and Company, 4^{th} edition.

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

Taylor, W. R. (1984). An algorithm to compare secondary structure predictions. *Journal of Molecular Biology*, 173:512–514.

Tomb, J.-F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Doughtery, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzegerald, L. M., Lee, N., Adams, M. D., Hickey, E. K., Berg, D. E., Gocayne, J. D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M. D., Weidman, J. M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S., Borodovsky, M., Karp, P. D., Smith, H. O., Fraser, C. M., and Venter, J. C. (1997). The complete genome sequence of the gastric pathogen *heliobacter pylori*. *Nature*, 388:539–547.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., and et al (2001). The sequence of the human genome. *Science*, 291:1304–1351.

Watson, J. D. and Crick, F. H. C. (1953). A structure for deoxyribose nucleic acid. *Nature*, 171:737–739.

White, J. V., Stultz, C. M., and Smith, T. F. (1994). Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. *Mathematical Biosciences*, 119:35–75.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

Wilmot, C. M. and Thornton, J. M. (1988). Analysis and prediction of the different types of β-turn in proteins. *Journal of Molecular Biology*, 203:221–232.

Wilson, C. and Doniach, S. (1989). A computer model to dynamically simulate protein folding: Studies with crambin. *Proteins: Structure, Function, and Genetics*, 6:193–209.

Wouters, M. A. and Curmi, P. M. G. (1995). An analysis of side chain interactions and pair correlations within antiparallel β-sheets: The differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs. *Proteins: Structure, Function, and Genetics*, 22:119–131.

Yi, T. M. and Lander, E. S. (1993). Protein secondary structure prediction using nearest-neighbor methods. *Journal of Molecular Biology*, 232:1117–1129.

Zhou, H. X., Lyu, P., Wemmer, D. E., and Kallenbach, N. R. (1994). α-helix capping in synthetic model peptides by reciprocal side chain-main chain interactions: Evidence for an N-terminal "capping box". *Proteins: Structure, Function, and Genetics*, 18:1–7.

Zhu, H. and Braun, W. (1999). Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of β-sheet formation in proteins. *Protein Science*, 8:326–342.

Zimmermann, K. (1994). When awaiting 'Bio' Champollion: Dynamic programming regularization of the protein secondary structure predictions. *Protein Engineering*, 7:1197–1202.

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. *Bulletin of Mathematical Biology*, 46(4):591–621.

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