29

Predicting Protein Secondary and Supersecondary Structure

Mona Singh
Princeton University

29.1 Introduction

Proteins play a key role in almost all biological processes. They take part in, for example, maintaining the structural integrity of the cell, transport and storage of small molecules, catalysis, regulation, signaling and the immune system. Linear protein molecules fold up into specific three-dimensional structures, and their functional properties depend intricately upon their structures. As a result, there has been much effort, both experimental and computational, in determining protein structures.

Protein structures are determined experimentally using either x-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. While both methods are increasingly being applied in a high-throughput manner, structure determination is not yet a straightforward process. X-ray crystallography is limited by the difficulty of getting some proteins to form crystals, and NMR can only be applied to relatively small protein molecules. As a result, whereas whole-genome sequencing efforts have led to large numbers of known protein sequences, their corresponding protein structures are being determined at a significantly slower pace. On the other hand, despite decades of work, the problem of predicting the full three-dimensional structure of a protein from its sequence remains unsolved. Nevertheless, computational methods can provide a first step in protein structure determination, and sequence-based methods are routinely used to help characterize protein structure. In this chapter, we review some of the computational methods developed for predicting local

0-8493-8597-0/01/$0.00+$1.50

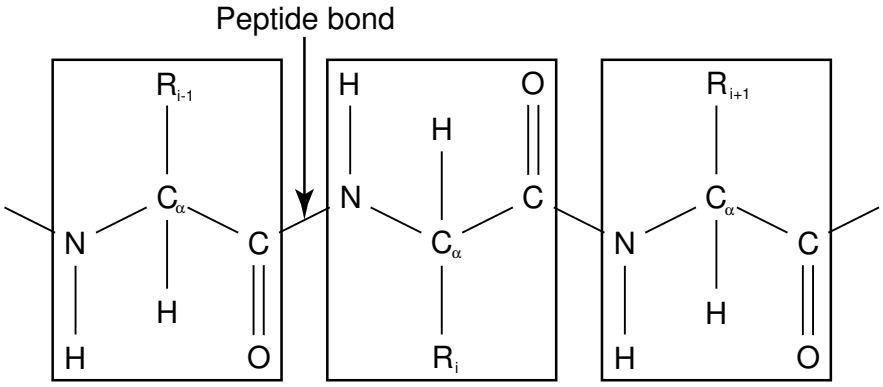

FIGURE 29.1: Proteins are polymers of amino acids. Each amino acid has the same fundamental structure (boxed), differing only in the atoms making up the side chain. Here, the i-th side chain in the protein sequence is designated by R_i. The carbon atom to which the amino group, carboxyl group, and side chain are attached is called the alpha carbon (C_α). Two amino acids $i-1$ and i are linked linearly through a peptide bond between the carboxyl group of amino acid $i-1$ and the amino group of amino acid i; a water molecule is removed in the process of bond formation.

aspects of protein structure.

29.1.1 Background

We begin by giving some introductory background to protein structure; there are many excellent sources for further information (e.g.,[BT99, Les01, Ric81]).

A protein molecule is formed from a chain of amino acids. Each amino acid consists of a central carbon atom (C_α), and attached to this carbon are a hydrogen atom, an amino group (NH_2), a carboxyl group (COOH) and a *side chain* that characterizes the amino acid. The amino acids of a protein are connected in sequence with the carboxyl group of one amino acid forming a peptide bond with the amino group of the next amino acid (Figure 29.1). Successive bonds make up the protein backbone, and the repeating amino-acid units (also called residues) within the protein consist of both the main-chain atoms that comprise the backbone as well as the side-chain atoms.

There are 20 side chains specified by the genetic code, and each is referred to by a one-letter code. A protein sequence can thus be described by a string over a 20-letter alphabet, and the *primary structure* of a protein refers to the covalent structure specified by its sequence (i.e., Figure 29.1), along with its disulfide bonds. The 20 side chains vary in atomic composition, and thus have different chemical properties. Some side chains are non-polar, or hydrophobic, because of their unfavorable interactions with water. Side chains have many other characteristics, and different side chains are commonly described as being positively charged, negatively charged, polar, small or large. Hydrophobic amino acids include isoleucine (I), leucine (L), methionine (M), phenylalanine (F) and valine (V). Arginine (R) and lysine (K) are positively charged in physiological pH, and aspartic acid (D) and glutamic acid (D) are negatively charged. Polar amino acids include asparagine (N), glutamine (Q) histidine (H), serine (S) and threonine (T). Alanine (A) is a small amino acid that is non-polar. Glycine (G) is the smallest amino acid, with just a hydrogen. Cysteine (C) can take part in disulfide bridges. Proline (P) has the strongest stereochemical constraints, and

tryptophan (W) and tyrosine (Y) are large, ring-shaped amino acids. There are many other (and sometimes conflicting) ways to classify and describe the amino acids.

The differences in physico-chemical properties of side chains result in the diversity of three-dimensional protein folds observed in nature. In particular, each possible structural conformation brings together a different set of amino acids, and the energy of the conformation is determined by the interactions of the side-chain and main-chain atoms with each other, as well as with solvent and ligands. There are many forces driving protein folding; for water-soluble proteins, the most dominant is the hydrophobic effect, or the tendency of hydrophobic amino acids to avoid water and bury themselves within the core of the protein. Hydrogen bonding, electrostatic interactions and van der Waals forces are also very important.

From a structural perspective, it is useful to think of protein chains as subdivided into peptide units consisting of the main-chain atoms between successive C_α atoms. In protein structures, the atoms in a peptide unit are fixed in a plane with bond lengths and angles similar in all units. Each peptide unit essentially has only two degrees of freedom, given by rotations around its N-C_α and C_α-C bonds. Phi (ϕ) refers to the angle of rotation around the N-C_α bond, and psi (ψ) refers to the angle of rotation around the C_α-C bond. The entire backbone conformation of a protein can thus be specified with a series of ϕ and ψ angles. Only certain combinations of ϕ and ψ angles are observed in protein backbones, due to steric constraints between main-chain and side-chain atoms.

As a result of the hydrophobic effect, the interior of water-soluble proteins form a hydrophobic core. However, a protein backbone is highly polar, and this is unfavorable in the hydrophobic core environment; these main-chain polar groups can be neutralized via the formation of hydrogen bonds. *Secondary structure* is the "local" ordered structure brought about via hydrogen bonding mainly within the backbone. Regular secondary structures include *α-helices* and *β-sheets* (Figure 29.2). A canonical α-helix has 3.6 residues per turn, and is built up from a contiguous amino acid segment via backbone-backbone hydrogen bond formation between amino acids in positions i and $i + 4$. The residues taking part in an α-helix have ϕ angles around $-60°$ and ψ angles around $-50°$. Alpha helices vary considerably in length, from four or five amino acids to several hundred as found in fibrous proteins. A β-strand is a more extended structure with 2.0 residues per turn. Values for ϕ and ψ vary, with typical values of $-140°$ and $130°$, respectively. A β-strand interacts via hydrogen bonds with other β-strands, which may be distant in sequence, to form a β-sheet. In parallel β-sheets, the strands run in one direction, whereas in antiparallel sheets, they run in alternating directions. In mixed sheets, some strands are parallel, and some are antiparallel. A β-strand is typically 5–10 residues in length, and on average, there are six strands per sheet. Coil or loop regions connect α-helices and β-sheets and have varying lengths and shapes.

Supersecondary structures, or *structural motifs*, are specific combinations of secondary structure elements, with specific geometric arrangements with respect to each other.[1] Common supersecondary motifs include α-helix hairpins, β hairpins, β-α-β motifs, and coiled coils. Elements of secondary structure and supersecondary structure can then combine to form the full three-dimensional fold of a protein, or its *tertiary structure*. Many proteins exist naturally as aggregates of two or more protein chains, and *quartenary structure* refers to the spatial arrangement of these protein subunits.

[1]Supersecondary structure is sometimes defined so as to require that the secondary structure units are consecutive in the protein sequence; we do not take that viewpoint here.

FIGURE 29.2: Schematic backbone conformations of an α-helix (left) and a β-sheet (right). An α-helix consists of contiguous amino acid residues. A β-sheet consists of individual β-strands, each of which is made up of contiguous amino acid residues. Here, a 5-stranded β-sheet, without the intervening regions, is shown.

29.1.2 Difficulty of general protein structure prediction

Experiments performed decades ago demonstrated that the information specifying the three-dimensional structure of a protein is contained in its amino acid sequence [AHSW61, Anf73], and it is generally believed that the native structure of the majority of proteins is the conformation that is thermodynamically most stable. It is now known that some proteins require specific proteins, or chaperones, to help them fold into their global free-energy minimum. A quantum mechanics treatment to predict structure is intractable for protein sequences, and thus physics-based methods for structure prediction typically use empirical molecular mechanics force fields. In these methods, the system is described as a set of potential energy terms (typically modeling bond lengths, bond angles, dihedral angles, van der Waals interactions and electrostatics), and the goal is to find, for any given protein sequence, the conformation that minimizes the potential energy function (e.g., see [BBO$^+$83]). The accuracy of state-of-the-art energy functions, the small energy differences between native and unfolded proteins, and the size of the conformational space that must be searched are all limiting factors in the overall performance of these physics-based methods. In the case where a protein is homologous to another with known structure, the search space is limited, as the homolog provides a template backbone; improved statistical methods for remote homology detection as well as the increasing number of solved protein structures have made such approaches more widely applicable. Purely statistical approaches have also been developed for predicting the tertiary structure of a protein. One such approach is known as *threading* [Sip90, BLE91, JTT92, BL93], where a sequence is aligned (or "threaded") onto all known backbones using an energy function that is estimated from observed amino acid frequencies in known protein structures. Many modern approaches use a combination of both statistics and physics; for example, in some of the more successful approaches for predicting protein structure, backbone fragments for particular subsequences are sampled from known structures, and then pieced together and evaluated using a molecular mechanics energy function [BCM$^+$03]. While there has been much progress in developing computational methods for predicting the three-dimensional structures of proteins, it is clear that the problem is far from being solved (e.g., [MFZH03, KWK$^+$03, ASHR03, TM03]).

29.1.3 A bottom-up approach

Because of the difficulty of the general protein structure prediction problem, an alternative approach for predicting protein structure is "bottom-up": here, the goal is to focus on specific, *local* three-dimensional structures, and develop specialized computational methods

for recognizing them within protein sequences. At the most basic level, a protein's secondary structure can be predicted. At the next level, computational methods may be developed to predict local supersecondary structures or structural motifs. Protein structure can also be characterized by identifying portions that are membrane-spanning, or by assessing the solvent accessibility of individual residues, though such subjects will not be reviewed here. By focusing on specific aspects of protein structures, it is possible to develop computational methods that can make high-confidence predictions; these can then be used to constrain methods that attempt to predict tertiary structure. At the same time, one hope is that ultimately it will be possible to build up a "library" of increasingly complex structures that can be recognized via specialized computational methods, and that this library may provide an alternative means for predicting the tertiary structures of proteins.

In the remaining portion of this chapter, we review computational techniques that have been developed for predicting secondary and supersecondary structures. While the most accurate predictions of structure are made by detecting homology to proteins with known structure, we primarily focus on methods that can make predictions even if there are no such homologs. Since there have been hundreds of papers written on predicting the secondary and supersecondary structure of proteins, we will only have a chance to discuss a small subset of the many important papers in the field.

29.2 Secondary structure

Most commonly, the secondary structure prediction problem is formulated as follows: given a protein sequence with amino acids $r_1 r_2 \ldots r_n$, predict whether each amino acid r_i is in an $\alpha-$helix (**H**), a $\beta-$strand (**E**), or neither (**C**). Predictions of secondary structure are typically judged via the *3-state* accuracy (Q_3), which is the percent of residues for which a method's predicted secondary structure (**H**, **E**, or **C**) is correct. Since residues in known protein structures are approximately 30% in helices, 20% in strands and 50% in neither, a trivial algorithm that always predicts **C** has a 3-state accuracy of 50%. The 3-state accuracy measure does not convey many useful types of information. For example, it does not indicate whether one type of structure is predicted more successfully than another, whether some structure is over- or under- predicted, or whether errors are more likely along the boundaries of secondary structure units than within them. Nevertheless, 3-state accuracy is a concise, useful measure that is frequently used to compare how well different methods perform. Other methods to judge the quality of secondary structure predictions include the Matthews correlation coefficient [Mat75] and measures of how well the predicted secondary structure segments overlap the actual ones [RSS94b, ZVFR99].

Secondary structural elements are readily evident in the crystal structures of proteins, and are defined operationally based primarily on their hydrogen bonding patterns. Given the 3D atomic coordinates of a protein structure, there are several automated means for extracting secondary structure, including DSSP [KS83a] and STRIDE [FA95]. The assignment of secondary structure to each amino acid is not completely well-defined, and these two programs differ on approximately 5% of residues (e.g.,see [CB99b]). Both DSSP and STRIDE report detailed descriptions of secondary structure. For example, the DSSP method has eight secondary structure classifications: **H**, α-helix; **E**, β-strand; **G**, 3_{10} helix, a helix with backbone-backbone hydrogen bonds between positions i and $i+3$; **I**, π-helix, a helix with backbone-backbone hydrogen bonds between positions i and $i+5$; **B**, bridge, a single residue β-strand; **T**, a hydrogen bonded turn; **S**, bend; and **C**, any residue that does not belong to any of the previous seven groups.

There are different schemes for translating the more detailed descriptions given by DSSP

and STRIDE into the three broad categories corresponding to helix, sheet and other. One scheme translates all helices (**H**, **G**, and **I**) into **H**, bridges and strands (**E**, **B**) into **E** and every thing else (**T**, **S**, **C**) into **C**. An alternative scheme takes the DSSP categories of **H** and **E** as helix and strand, and maps all other categories into **C**. The reported performance of a secondary structure prediction method can vary depending on which precise translation scheme is used, with the second scheme leading to higher estimates of accuracy [CB99b].

Testing of secondary structure prediction methods has improved over the years. We note that whereas the PDB (the Protein Data Bank of solved structures [BWF$^+$00]) contains structures for many very similar sequences, the training set used for estimating parameters should not contain sequences that are too similar to those in the test set. In particular, a protein sequence in the test set should be less than 25–30% similar to any sequence in the training set. Otherwise, reported accuracy is likely to be an overestimate of actual accuracy. Methods are typically tested using N-fold cross-validation, where a dataset is split into N parts. Each part is in turn left out of the training set and performance is judged on it. The performance of the method is the average performance over each left out part.

Early secondary structure prediction methods (such as Chou-Fasman and GOR, described below) have a 3-state cross-validation accuracy of 50–60%. Today's methods have an accuracy of $> 75\%$.

29.2.1 Early approaches

The earliest approaches for secondary structure prediction considered just single amino acid statistics and properties, and were limited by the small number of proteins with solved structures. While these early methods are not state-of-the-art, they are natural first attempts to the secondary structure prediction problem, and are the basis of many subsequent approaches. Below, we consider three of the most well-known early secondary structure prediction methods.

Chou-Fasman method. One of the first approaches for predicting protein secondary structure, due to Chou and Fasman [CF74], uses a combination of statistical and heuristic rules. First, using a set of solved protein structures, "propensities" are calculated for each amino acid a_i in each structural conformation s_j, by taking the frequency of a_i in each structural conformation, and then normalizing by the frequency of this amino acid in all structural conformations. That is, if a residue is drawn at random from the space of protein sequences, and its amino acid identity A and structural class S are considered, propensities are computed as $\Pr(A = a_i|S = s_j)/\Pr(A = a_i)$.[2] These propensities capture the most basic concept in predicting protein secondary structure: different amino acids occur preferentially in different secondary structure elements.

Once the propensities are calculated, they are used to categorize each amino acid as either a helix-former, a helix-breaker, or helix-indifferent. Each amino acid is also categorized as either a sheet-former, a sheet-breaker, or sheet-indifferent. For example, as expected, glycine and proline have low helical propensities and are thus categorized as helix-breakers. Then, when a sequence is input, "nucleation sites" are identified as short subsequences with a high-concentration of helix-formers (or sheet-formers). These sites are found with heuristic

[2]Sometimes propensities are defined by considering the frequency of a particular structural conformation given an amino acid, and normalizing by the frequency of that structural conformation. These two formulations are equivalent since $\Pr(A = a_i|S = s_j)/\Pr(A = a_i) = \Pr(S = s_j|A = a_i)/\Pr(S = s_j)$.

rules (e.g., "a sequence of six amino acids with at least four helix-formers, and no helix-breakers"), and then extended by adding residues at each end, while maintaining an average propensity greater than some threshold. Finally, overlaps between conflicting predictions are resolved using heuristic rules.

GOR method. The GOR method [GOR78] formalizes the secondary structure prediction problem within an information-theoretic framework. If x and y are any two events, the definition of the information that y carries on the occurrence of event x is [Fan61]:

$$I(x;y) = \log\left(\frac{\Pr(x|y)}{\Pr(x)}\right). \tag{29.1}$$

For the task at hand, the goal is to predict the the structural conformation S_j of residue R_j in a protein sequence, and the GOR method estimates the information that the surrounding "local" 17-long window contains about it:

$$I(S_j; R_{j-8}, \ldots, R_j, \ldots R_{j+8}) \quad = \quad \log\left(\frac{\Pr(S_j|R_{j-8}, \ldots R_j, \ldots, R_{j+8})}{\Pr(S_j)}\right). \tag{29.2}$$

In fact, each structural class x is considered in turn, and the following value, representing the preference for x over all other alternatives $\overline{x}$ is computed:

$$I(S_j = x : \overline{x}; R_{j-8}, \ldots, R_j, \ldots R_{j+8}) = I(S_j = x; R_{j-8}, \ldots, R_j, \ldots, R_{j+8}) - I(S_j = \overline{x}; R_{j-8}, \ldots R_j, \ldots, R_{j+8}).$$

To predict residue R_j's structural conformation, these values are computed for all structural states, and the one that has the highest value is taken as the prediction.

Because there are far too many possible sequences of length 17, it is not possible to estimate $\Pr(S_j|R_{j-8}, \ldots R_j, \ldots, R_{j+8})$ with any reliability. Instead, the original GOR method assumes that the values of interest can be estimated using single residue statistics:

$$I(S_j = x; R_{j-8}, \ldots, R_j, \ldots, R_{j+8}) = \sum_{m=-8}^{m=8} I(S_j = x; R_{j+m}), \tag{29.3}$$

where by definition $I(S_j = x; R_{j+m}) = \log(\Pr(S_j = x|R_{j+m})/\Pr(S_j = x))$.[3] $I(S_j = x; R_{j+m})$ represents the information carried by a residue at position $j+m$ on the conformation assumed by the residue at j. If $m \neq 0$, this does not take into account the type of residue at position j, and the intuition is that it describes the interaction of the side chain of residue $j+m$ with the backbone of residue j. For each structural class, this method requires estimating 20×17 parameters.

Lim method. A complicated, stereochemical rule-based approach for predicting secondary structure in globular proteins was developed at about the same time as the statistical methods discussed above. In this method, longer-range interactions between residues are considered. If the protein sequence is $r_1 r_2 \ldots r_n$, then for the i-th residue, the following pairs and triples are considered particularly important for helical regions: (r_i, r_{i+1}), (r_i, r_{i+3}), (r_i, r_{i+4}), (r_i, r_{i+1}, r_{i+4}), (r_i, r_{i+3}, r_{i+4}). Note that residues three and four apart are considered, as they lie on the same face of an α-helix. Similarly, the pair (r_i, r_{i+2}) contains

[3]Note that when $m = 0$, these values are equivalent to taking the log of the Chou-Fasman propensity values.

residues on the same face of a β-strand. Pairs and triplets of particular amino acids are then deemed as compatible or incompatible with helices and strands based on various rules that try to ensure that these residues present a face that allows tight packing of hydrophobic cores. Factors used to determine these rules include each amino acid's size, hydrophobicity, charge, and its ability to form hydrogen bonds. For example, if a protein sequence has hydrophobic residues every three to four residues, this method predicts compatibility with an α-helix, as this would result in one side of the helix being hydrophobic, thus facilitating packing onto the rest of the protein structure.

29.2.2 Incorporating local dependencies

Whereas the first statistical methods for predicting protein secondary structure examined each amino acid individually, later approaches began to consider higher-order residue interactions, either within statistical approaches or via machine learning methods. Reported 3-state accuracies for most of these methods are above 60%.

Information theory approaches. One approach to incorporate higher-order residue interactions is an extension to the original GOR method [GGR87]. The notion of conditional information is helpful here. In particular, $I(x; y_2|y_1)$ is defined as $\log(\Pr(x|y_1, y_2)/P(x|y_1))$. Note that $I(x; y_1, y_2, \ldots, y_n) = I(x; y_1) + I(x; y_2|y_1) + \ldots + I(x; y_n|y_1, y_2, \ldots y_{n-1})$. Instead of the assumption made in equation 29.3, the following assumption is made:

$$I(S_j = x; R_{j-8}, \ldots, R_j, \ldots, R_{j+8}) = I(S_j = x; R_j) + \sum_{m=-8, m\neq 0}^{m=8} I(S_j = x; R_{j+m}|R_j).$$

This formulation incorporates the information carried by the residue at $j+m$ on the conformation of the residue at j, taking into account the type of residue at position j. Note that by changing these assumptions, different pairwise or higher-order residue interactions may be considered. Later versions of the GOR algorithm do precisely this (e.g., see [KTJG02]).

Nearest-neighbor approaches. Nearest-neighbor methods classify test instances according to the classifications of "nearby" training examples. In the context of secondary structure prediction, the overall approach is to predict the secondary structure of a residue in a protein sequence by considering a window of residues surrounding it, and finding similar sequence segments in proteins of known structure. The assumption is that short, very similar sequences of amino acids have similar secondary structure even if they come from non-homologous proteins. The known secondary structures of the middle residue in each of these segments are then combined to make a prediction, either via a simple voting scheme or a weighted voting scheme, with segments more similar to the target segment weighed more. Early nearest-neighbor approaches include [NO86, LRG86]. Similar segments can be found via sequence similarity, or via structural profiles [BLE91], as in [LL93].

Neural network approaches. Neural networks provide another means for capturing higher-order residue interactions. They were first applied to predict secondary structure by [QS88, HK89], and some of the most successful modern methods are also based on neural networks (e.g., [RS93] and its successors).

Because neural nets are widely used in the field of secondary structure prediction, we briefly describe them here. Neural networks, loosely based on biological neurons, are machine learning methods that learn to classify input vectors into two or more categories. Feedforward neural networks consist of two or more connected layers. The first layer is the input layer, and the last layer is the output layer that indicates the predicted category of

the input. All other layers are called hidden layers. A simple neural network with no hidden units is given in Figure 29.3. The inputs can be encapsulated in a vector $\vec{x} = (x_1, \ldots, x_m)^T$, and each of the input edges has a corresponding weight, giving $\vec{w} = (w_1, \ldots, w_m)^T$. Each input is multiplied by the corresponding weight of its edge. Then, the network computes a weighted sum, and feeds it into some activation or continuous threshold function σ. For example, $\sigma(a)$ could be $\frac{1}{1+e^{-a}}$, which is a sigmoidal function with values between 0 and 1.[4] Thus, the function computed by this simple neural network is given by $\sigma(\vec{w} \cdot \vec{x})$, and is essentially linear. In most cases, a neural net must learn the weights from a training set of input vectors $\{\vec{x_i}\}$ where the target value t_i for each is known. For example, in the scenario described, there may be two classes of examples with target values of 0 and 1. Typically, the goal is to find the weights $\vec{w}$ minimizing some error function (e.g., the squared error $E = \sum_i (\sigma(\vec{w} \cdot \vec{x_i}) - t_i)^2$). Such a $\vec{w}$ can be found via gradient descent.[5] A full-blown neural net is built up from a set of simpler units that are interconnected in some topology so that the outputs of some units become the inputs of other units (e.g., see Figure 29.4). The gradient descent procedure for arbitrary neural networks is implemented via the back-propagation algorithm [RHW86b, RHW86a]. While neural nets with multiple layers are not as easy to interpret as those without hidden layers, they can approximate any continuous function $f : \mathcal{R}^m \rightarrow \mathcal{R}$ as long as they have a sufficient number of hidden units and at least two hidden layers [Cyb89].

The two early neural-network approaches to secondary structure prediction use similar neural network topologies. Holley and Karplus [HK89] build a neural net that tries to predict the secondary structure of a residue r_j by considering residues $r_{j-8}, \ldots, r_j, \ldots, r_{j+8}$. Each of these residues is represented with 21 bits, corresponding to the 20 amino acids and an extra bit for the case where the window overlaps the beginning or end of the sequence. Thus, each example is represented by 17×21 bits, with only 17 non-zero entries. The topology of the neural net has one hidden layer with two nodes, and two output nodes, one corresponding to helix and the other to sheet (see Figure 29.4). For the training process, if the middle residue's secondary structure is helix, then the target output has the helix output set to 1 and the sheet output set to 0. For a new sequence, helix is assigned to four or more adjacent residues each with helix output value greater than both the sheet output value and some threshold. Strand is assigned similarly, though only two adjacent strand residues are required.

Qian and Sejnowski [QS88] have a slightly different network topology. They consider 13-long windows, and have three output units (one for each of the three states). More significantly, they additionally use a cascade of neural networks in order to capture correlations between secondary structure assignments of neighboring residues. In particular, they show improved performance by first training a neural net to predict the secondary structure of a central amino acid, and then taking the outputs for adjacent residues using this trained

[4] While a strict 0/1 threshold function can also be used, a continuous function is preferred for ease of optimization.

[5] There are many other approaches to find a set of weights that "best" linearly separate two classes. For example, the support vector machine framework (SVM) [Vap98] finds weights so that the margin between the two classes of examples is maximized; that is, an SVM finds the weights by maximizing the distance between the hyperplane specified by the weights and the closest training examples. In the case where the two classes are not linearly separable, the data are typically embedded in a higher dimensional space where they are separable. An alternative approach, linear discriminant analysis, tries to find a set of weights so that when considering $D_x = \vec{w} \cdot \vec{x}$ for all examples $\vec{x}$, these values are as close as possible within the same class and as far apart as possible between classes.

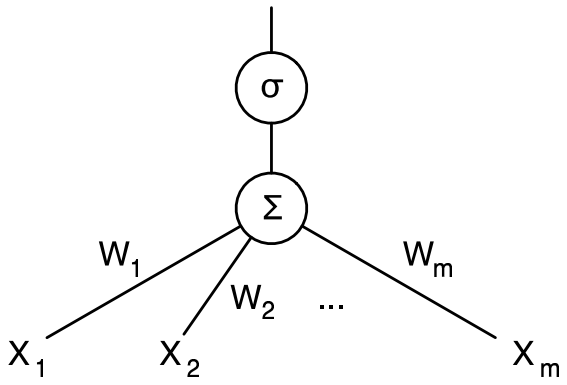

FIGURE 29.3: A simple neural network with no hidden units. There are m inputs $x_1 \ldots x_m$, and the neural net computes a function on these inputs by first calculating $\sum_i w_i x_i$, and then using this as input to an activation function σ.

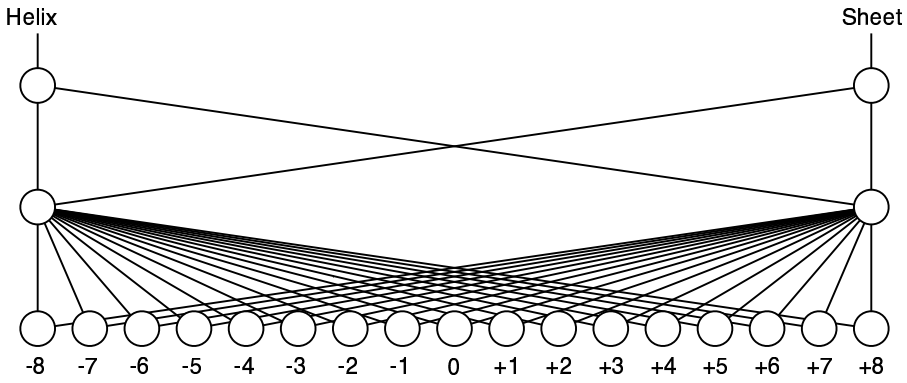

FIGURE 29.4: A neural network topology for predicting secondary structure [HK89]. To predict the secondary structure of the middle residue, eight residues on either side are considered. Each of the 17 input units drawn actually consists of 21 mutually exclusive binary inputs, one for each possible amino acid, and one used when the window overlaps the end of the protein sequence. There is one hidden layer with two units, and an output layer with two units. Here, each node in the hidden layer and output layer contains both a summation and activation component. The basic neural network of [QS88] is similar, but with 13 input units, more hidden units, and a third output unit corresponding to coil.

network and feeding them into a second network. The input layer of this second network has 13 groups, with three units per group, one for each output unit from the first network.

29.2.3 Exploiting evolutionary information

It is well-known that protein structure is more conserved than protein sequence, and that two sequences that share more than 30% sequence identity are likely to have similar structures. Thus, when predicting the secondary structure of a particular protein sequence, predictions for its homologs may also prove useful. Additionally, conservation evident in multiple

sequence alignments (MSAs) of homologs helps reveal which amino acids are likely to be functionally or structurally important, and may highlight the characteristic hydrophobic patternings of secondary structure elements. For example, surface-exposed loop regions that are not important functionally tend to be part of variable regions in MSAs.

A natural first attempt to use homologous proteins in order to improve secondary structure prediction might make predictions for each homolog, and then average (or otherwise combine) these predictions for corresponding amino acids [ZBTS87, KTJG02]. Alternatively, information from all sequences may be used at once in order to make one set of predictions. This is the approach taken by Rost and Sander [RS93], and their neural network based program was the first to surpass 70% 3-state accuracy. The use of evolutionary information is critical for the improved performance, and all modern approaches use evolutionary information in making secondary structure predictions.

To make predictions about a single protein sequence, the approach of [RS93] begins with homologs gathered via database search. These homologs are then aligned in a MSA, and a profile is made. In particular, for each column j in the MSA, the frequency of each amino acid i in the column is computed. To determine the secondary structure of residue r_j, a sequence-to-structure neural network considers a window of 13 residues $r_{j-6} \ldots r_j \ldots r_{j+6}$. For each residue in the window, instead of giving just the identity of this residue as input to the neural net, the frequencies of all amino acids in the corresponding column of the MSA are fed into the neural network. These frequencies encapsulate the evolutionary constraints on each residue. Other features, including overall amino acid composition, are also input to the network. This network has three output units, one for each secondary structure state. Similar to [QS88], the output of this first level neural network is fed to a structure-to-structure network. Finally, a jury system is used to make the final predictions. Because neural networks are sensitive to topology, the set of training data, the order of training, as well as other parameters, several networks are trained while varying these parameters. The jury system level takes as input the results from each of these nets and averages them. The secondary structure with the highest average score is output as the prediction. A very useful feature of this approach is a per-position reliability index, where higher numbers correspond to more confident predictions. Neural nets have become the most common approach to secondary structure prediction; more recent extensions have included the use of recurrent neural nets to capture non-local interactions [BBF$^+$99, PPRB02].

Incorporating evolutionary information into other basic secondary structure prediction methods also results in improved performance, and while it was initially suggested otherwise, it is unlikely that there is some special feature of neural nets that makes them particularly well-suited to predicting secondary structure. For example, similar performance has also been achieved using support vector machines (SVMs) [HS01]. Adding evolutionary information to nearest-neighbor approaches [SS95] also performs competitively; here, predictions are made individually for each homolog and then combined. Another MSA approach with similar reported performance [KS96] uses linear discriminant analysis to combine several predictive attributes. These include: residue propensities, computed as in GOR [GOR78]; distance from the end of the protein sequence; moments of hydrophobicity [EWT84] for each residue under the assumption that it and its three neighboring residues in each direction are in either helices or sheets; whether or not an insertion or deletion is observed in any of the homologs in the MSA; and an entropy-based measurement of residue conservation. Sequence correlations are captured by feeding in the output of the first linear discrimination function into another one, and additionally incorporating smoothing of features over nearby residues, predicted ratios of α-helix and β-strand, and measures of sequence amino acid content.

While most methods incorporate evolutionary information using global MSAs, an alter-

nate method relies solely on pairwise local alignments [FA96]. A weight is computed for each pairwise alignment based on its score and length. For each residue in the original sequence, the weighted sum over all aligned sequences is computed independently for several propensity values, which are then combined using a rule-based system to make a final prediction. These propensity values are interesting, as several of them try to incorporate non-local interactions [FA96]. In particular, β-strand hydrogen bonding parallel and anti-parallel propensities (obtained from known structures) are computed between neighboring sequence fragments, and helical hydrogen bonding propensities are computed for fragments by considering residues i and $i+4$. A propensity concerning β-turn is also used ([HT94], see below), as well as helical, strand and coil propensities computed using a nearest-neighbor approach.

29.2.4 Recent developments and conclusions

Further improvements in performance have come from better remote homology detection (e.g., using PSI-BLAST [AMS$^+$97] or hidden Markov models [KBH98, KKD$^+$03]), and larger sequence databases [CB99a, Ros01]. For example, Jones [Jon99] obtained better performance than [RS93] ($>$ 75% 3-state accuracy) using a similar neural network architecture (without a jury system layer), but where homologs are first detected via PSI-BLAST [AMS$^+$97]. PSI-BLAST is an iterative database searching method that uses homologs found in one iteration to build a profile used for searching in the next iteration. The detected homologs are then input into the neural network via the profile provided by PSI-BLAST; this profile incorporates sequence weighting so that several closely-related homologs detected in the database do not overwhelm the contribution of more remote homologs. It is likely that sequence weighting also plays a role in the improved performance of this method, as it has been shown that predictions improve when getting rid of closely related homologs [CB99b].

Several authors have also attempted to predict secondary structure by combining the results of several different programs. For example, Cuff and Barton [CB99b] predict secondary structure by taking the most commonly predicted state by four methods [RS93, SS95, KS96, FA97], and show a modest improvement in performance. Existing approaches have also been combined using machine-learning methods such as linear discriminant analysis, decision trees and neural nets, and have shown to give upto a 3% improvement in 3-state accuracy over the best individual method [KOS$^+$00].

Future evaluation. An important recent development has been to set up continuous evaluation procedures (such as EVA [EMRP$^+$01]). Protein sequences with newly determined structures are sent to the webservers of the programs being evaluated. In general, evaluation and comparison of methods is often difficult, due to differences in the evaluation methodology and the changing structural databases; thus, a community-wide approach such as this should have great impact on future development of secondary structure prediction methods.

Limitations of secondary structure prediction. In general, it is believed that α-helices are easier to predict than β-sheets. A recent evaluation found that helices were predicted 9.5% more accurately than strands [ASHR03]. This may be because the hydrogen bonding patterns for α-helices are among amino acids in close proximity to each other, and those for β-sheets are not. Additionally, shorter secondary structure elements are harder to predict, presumably because the signal is not strong enough from these fragments.

Clearly, protein secondary structure is influenced by both short- and long-range interactions. It has been demonstrated that there are 11-long amino acid sequences that can fold into an α-helix in one context, and a β-sheet in another [MK96]. However, even as-

suming that long-range tertiary interactions can be incorporated into secondary structure prediction algorithms, the best possible 3-state accuracy will not be 100%. First, assignment of secondary structure is not always clear even when there is a crystal structure. This is evident from the observed differences between STRIDE and DSSP [CB99b]. Additionally, while secondary structure predictions improve when incorporating evolutionary information, homologous structures do not share identical descriptions of secondary structure assignments [RSS94a]. Even when a query sequence can be aligned confidently to a sequence of known structure, the alignment will produce a secondary structure "prediction" with 3-state accuracy of only 88% on average [RSS94a]. Accordingly, while secondary structure prediction methods continue to improve, it is unlikely that any method that does not also solve the tertiary structure prediction problem will achieve ideal performance in predicting secondary structure.

29.3 Tight turns

Tight turns are secondary structure elements consisting of short backbone fragments (no more than six residues) where the backbone reverses its overall direction. Tight turns allow a protein to fold into a compact globular structure, and identifying them correctly in a protein sequence limits the search space of possible folds for the sequence. Tight turns are also important because they are often on the surface of proteins, and thus may play a role in molecular interactions. Tight turns are categorized according to their lengths into $\delta-$, $\gamma-$, $\beta-$, $\alpha-$ and $\pi-$ turns, which consist of two, three, four, five, and six residues respectively.

Computational methods have been developed for recognizing tight turns in protein structures, with most of the work focusing on β-turns, which occur most frequently in protein structures. Approximately one-quarter of all protein residues are in β-turns [KS83a]. A β-turn is defined as four consecutive residues r_i, r_{i+1}, r_{i+2} and r_{i+3}, where the distance between the C_α of residue r_i and the C_α of residue r_{i+3} is < 7 Å, and the central two residues are not helical. These β-turns can be further assigned to one of several (6–10) classes on the basis of the backbone ϕ and ψ angles of residues r_{i+1} and r_{i+2} [Ven68, LMS73, Ric81, HT94, Cho00]. The first methods for predicting β-turns focused on identifying which residues take part in β-turns [LMS71, CF79], and later methods have additionally attempted to predict the type of β-turn [WT88]. Some β-turn types show preferences for particular topological environments; for example, type I′ and type II′ β-turns are preferentially found in β hairpins [ST85].

As with 3-state secondary structure prediction, methods to predict β-turns fall into two classes: probabilistic methods and machine-learning methods. The earliest probabilistic methods computed the probability that a certain amino acid a_i is located at the j-th position in a β-turn by dividing the number of times the amino acid a_i occurred in the j-th position of a turn by the total occurrences of amino acid a_i [LMS71]. Assuming independence between positions, the probability that a certain 4-long window is an occurrence of a β-turn is calculated by the product of the appropriate four terms, and a cutoff for prediction is chosen. These predictions can be further refined so that a 4-long window that has helical or sheet propensity that is larger than its β-turn propensity is eliminated [CF79]; structural propensities are defined as in [CF74]. Modifications of this basic approach to predict turn types include [WT88, WT90, HT94].

Other probabilistic methods consider each possibility Ψ (where Ψ can be each type of β-turn as well as non-β-turns) in turn, and compute the probability of observing a particular 4-long window given that it is an instance of Ψ. In particular, given a subsequence $r_1r_2r_3r_4$,

it is scored by considered a random subsequence $R_1R_2R_3R_4$ and computing

$$\Pr(R_1 = r_1, R_2 = r_2, R_3 = r_3, R_4 = r_4|\Psi).$$

The possibility Ψ giving the largest value is taken as the prediction. Assuming that each position is independent of every other, this simplifies to

$$\prod_{i=1}^{i=4} \Pr(R_i = r_i|\Psi).$$

For each type of β-turn, probabilities are estimated from known structures for each of the four positions. Later models [ZC97] consider the spatial arrangement of β-turns and assumed dependencies between the first and fourth position, and the second and third positions:

$$\Pr(R_1 = r_1|\Psi)\Pr(R_2 = r_2|\Psi)\Pr(R_3 = r_3|R_2 = r_2, \Psi)\Pr(R_4 = r_4|R_1 = r_1, \Psi).$$

Alternate models make the 1st order Markov assumption that all dependencies can be captured by considering adjacent residues [Cho97, CB97]:

$$\Pr(R_1 = r_1|\Psi)\Pr(R_2 = r_2|R_1 = r_1, \Psi)\Pr(R_3 = r_3|R_2 = r_2, \Psi)\Pr(R_4 = r_4|R_3 = r_3, \Psi).$$

The earliest neural network approaches [MFS89] to β-turn prediction take as input a 4-long window of amino acids (each residue is represented with 20 bits), and include a hidden layer. There are four output nodes, two for the most common $\beta-$turn classes, one for all other β-turns, and one for non-β-turns. Later approaches subdivide the problem into first predicting whether a window contains a β-turn and then predicting the type of turn [SGT99]. As in neural network based approaches to predicting secondary structure [QS88, RS93], several layers of neural networks are used. In the first, a nine amino acid window is considered. Additionally, for each residue, secondary structure predictions (helix, sheet or other) are considered; inclusion of such predictions improves performance for both neural network [SGT99, KR03b] and statistical approaches [KR02] for β-turn prediction. The output for adjacent residues using this neural network are fed into a second structure-to-structure network, along with secondary structure predictions. Predictions are also filtered via a rule-based system. Finally, all data identified by the turn/not-turn networks as possibly taking part in β-turns are input to networks for turn types, with only 4-long amino acid windows considered. When several turn types can be potentially predicted for a particular window, the one with the largest score is taken as the prediction. As with 3-state secondary structure prediction, further improvements in β-turn prediction have been obtained by using evolutionary information, where each sequence position is encoded using a profile describing its amino acid distribution in a MSA [KR03b, KR04]. More recently, nearest-neighbor [Kim04] and SVMs [CLL$^+$03] have also been applied to predict β-turns.

Predictions of β-turns are not as reliable as 3-state predictions of secondary structure. Approximately 50% of β-turns can be identified with 75% of the sequence fragments predicted as β-turn actually being correct. Overall accuracy of predictions is around 75%; a method that always predicts non-β-turns would have similar accuracy. Furthermore, predictions of β-turn types are only possible for the most frequent turn types.

More recently, attempts have been made to predict γ-turns and α-turns [KR03a, CC99, CFLC03]. The computational techniques are very similar to the ones applied to β-turns. Perhaps due to the vastly fewer number of residues taking part in either γ- or α-turns, these methods have only had limited success.

29.4 Beta hairpins

Beta hairpins are one of the simplest supersecondary structures and are widespread in globular proteins. They consist of short loop regions (or turns) between antiparallel hydrogen bonded β-strands. Typically, the length of these loop regions is eight residues or less, with two residue loops being most common [ST85, KMB04]. Correct identification of such structures can significantly reduce the number of possible folds consistent with a given protein, as differing tertiary folds contain different arrangements and numbers of β-strands. As noted in [dlCHST02, KMB04], consecutive β-strands in a protein sequence can either form more "local" hairpin structures or "diverge" so that the β-strands may pair with other strands. Methods for predicting β hairpins have just begun to appear, and two recent approaches are based on neural networks.

In the first approach [dlCHST02], β hairpins are identified by first predicting secondary structure. Each predicted β-coil-β pattern is further evaluated by comparing it to all known β hairpins of the same length. Each comparison between the pattern and a known β hairpin results in 14 scores. These scores are computed based on the compatibility of the predicted secondary structures and solvent accessibilities with those known for the hairpin, and additionally incorporate the segment's turn potential, secondary structure elements' lengths, putative pairwise residue interactions, and pairwise residue contacts. These scores are then fed into a neural network that is trained to discriminate between hairpins and non-hairpins. Finally, all the database matches for a particular β-coil-β segment are evaluated, and if there are more than 10 predictions of a hairpin structure, the segment is predicted as a hairpin.

The second approach [KMB04] incorporates evolutionary information in predicting β hairpins. Homologs are obtained using PSI-BLAST [AMS$^+$97], and each position is represented via the underlying profile (as in [Jon99]). Two neural networks are trained, where the first predicts the state of the first residue in a turn, and the second predicts the state of the last residue of the turn. Each neural network predicts whether the residue being considered is the first (or last) residue of a hairpin, a diverging turn, or neither. To predict whether a residue is the start of turn, four residues before it and seven residues after it are considered. Similarly, to predict whether a residue is the end of turn, seven residues before it and four residues after it are considered. Thus, turns up to length eight are completely included in the input window. Each residue in the window is encoded using the appropriate column in the PSI-BLAST profile, as well as three additional parameters corresponding to secondary structure as predicted by [Jon99]. Finally, the per-residue predictions are combined to determine the probability of a particular structure (hairpin turn, other turn, or no turn) starting at residue i and ending at residue j. The authors additionally show that incorporating predictions of hairpins or diverging turns improves their method [SKHB97] for tertiary structure prediction.

The performance of the two approaches is not directly comparable, as the first considers hairpins of all lengths, and the second limits itself to hairpins with turn regions of length at most eight. It is likely that longer-range interactions are more difficult to predict. Additionally, the two approaches use different PDB training and testing sets, and report different fractions of β-coil-β patterns that are hairpins (40% vs. 60%). The approach of [dlCHST02] relies on the correct secondary structure prediction, and thus cannot predict β hairpins whose underlying secondary structure is not predicted correctly. Given an actual turn, the approach of [KMB04] identifies whether it is hairpin or diverging with accuracy 75.9%; a baseline performance of 60% is possible by predicting all turns as hairpin.

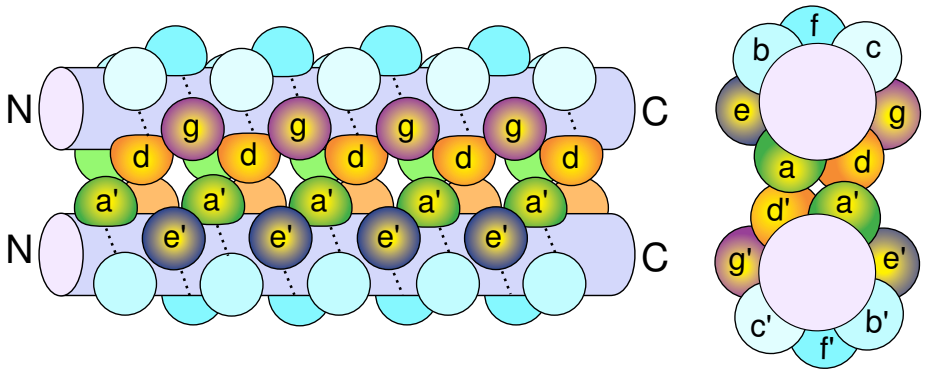

FIGURE 29.5: (a) Side view of a parallel 2-stranded coiled coil. (b) Top view of a parallel 2-stranded coiled coil. The interface between the α-helices in a coiled-coil structure is formed by residues at the core positions **a**, **d**, **e** and **g**. For notational convenience, positions in the two helices are distinguished by the prime notation (e.g., **a** and **a**$'$ are analogous positions in the two helices).

29.5 Coiled coils

The coiled coil is a ubiquitous protein structural motif that can mediate protein interactions. Roughly 5–7% of eukaryotic proteins contain coiled-coil regions. Coiled-coil structures are associated with several cellular functions, including transcription, oncogenesis, cell structure and membrane fusion. Coiled coils consist of two or more right-handed α-helices wrapped around each other with a slight left-handed superhelical twist. The helices in a coiled coil may associate with each other in a parallel or anti-parallel orientation, and the sequences making up the helices may either be the same (homo-oligomers) or different (hetero-oligomers). Helices taking part in coiled-coil structures exhibit a characteristic heptad repeat, denoted $(\mathbf{abcdefg})_\mathbf{n}$, spread out along two turns of the helix (see Figure 29.5). Residues at positions **a** and **d** tend to contain hydrophobic residues, and residues at positions **e** and **g** tend to contain charged or polar residues. The heptad repeat falls 20° short of two complete turns of a regular α-helix, and the supercoiling of the helices maintains that the **a** and **d** positions stay within the core of the structure. Coiled-coil helices pack with each other in a "knobs-into-hole" fashion [Cri53], where a residue in the **a** (or **d**) position is a "knob" that packs into a hole created by four residues on the other α-helix.

Just as secondary structure assignment from known three-dimensional structures is not unambiguous (e.g., [CB99b], and see above discussion), it is non-trivial to determine coiled coils in the set of solved structures. Different researchers may have different opinions on whether a particular structure is a coiled coil or not. The approach of [WW01] detects coiled coils by searching for knobs-into-holes packing. This approach identifies "true" coiled coils, as well as helical bundle domains where a subset of the helices interact with each other in a knobs-into-holes fashion.

Computational approaches have been developed both for identifying portions of protein sequences that can take part in coiled-coil structures, as well as for predicting specific interactions *between* coiled-coil proteins. While in principle it is possible to identify helices taking part in coiled coils by secondary structure prediction methods, in practice it is more effective to develop specialized methods for recognizing their hallmark heptad repeat. Most of the methods outlined below rely on having databases of known coiled-coil and non-coiled

coil sequences. Non-coiled coil databases can be derived from the PDB by excluding potential coiled-coil proteins. Coiled-coil databases are built both from analyzing the PDB, and from including fibrous proteins whose X-ray diffraction patterns reveal coiled-coil structures but do not permit high-resolution structure determination (review, [Coh98]).

29.5.1 Early approaches

The earliest approaches [Par82, LvDS91] to recognize coiled coils use sequences of known coiled-coil proteins, and construct a 20×7 table tabulating the frequency with which each amino acid is found in each of the seven heptad repeat positions, normalized by the frequency of the amino acid in all protein sequences. These values are very similar to the propensity values computed by the Chou and Fasman approach [CF74]. For example, for leucine and position **a**, the corresponding entry in the table is the percentage of position **a** residues in the coiled coil database which are leucine, divided by the percentage of residues in all protein sequences that are leucine. For each amino acid in a protein sequence, this approach considers all l-long windows that contain it.[6] Each of the l windows is considered with its first amino acid starting in each of the seven possible heptad repeat positions, and the heptad repeat proceeding uninterrupted in the window. Thus, $7l$ windows are considered for each residue, and each window is scored by taking the product of the propensities for each amino acid (in the appropriate heptad repeat position) in the window. The score for each residue is then the maximum score for any of the windows containing it, and the score for the sequence is the maximum score of any of its residues. Scores are converted to probabilities by approximating both the background and coiled-coil score distributions with Gaussians, and assuming that 1 in 30 residues is in a coiled coil.

This method has also been extended to recognize the "leucine zipper" family of coiled coils found in bZIP transcription factors. The bZIPs are a large family of eukaryotic transcription factors (review, [Hur95]), and their dimerization is mediated by the leucine zipper coiled-coil region. While the tendency is not uniformly true, leucine zippers tend to have leucines in the **d** position of the coiled coil. Early attempts to recognize leucine zippers focused on identifying leucine repeats, but since leucine is the most frequent amino acid, such patterns are frequently found by chance [BK89]. Both [HVSB96] and [BBRV98] find leucine zipper proteins by first identifying leucine repeats, and then requiring a coiled-coil prediction by [LvDS91]. [HVSB96] further uses both disallowed and highly preferred pairs of residues to identify leucine zipper coiled coils. The approach of [BBRV98] relaxes the requirement of a strict leucine repeat, and additionally focuses on identifying the short coiled-coil segments found in transcription factors.

29.5.2 Incorporating local dependencies

Subsequent approaches to predicting coiled-coil helices incorporate pairwise frequencies by explicitly considering the problem within a probabilistic framework [Ber95, BWW$^+$95]. This overall framework for coiled-coil prediction is similar to the information theory approaches described above for secondary structure prediction [GGR87]; however, the assumptions used in practice are very different. Here, the goal is to predict whether a subsequence

[6] A typical window length is 28 (four heptads), as it is thought that peptides that can form stable coiled coils in solution should be at least this length. Shorter windows can also be employed; typically, the discriminatory performance of methods deteriorate with shorter window sizes.

$z = r_1, r_2, \ldots, r_l$ is a coiled coil by estimating $\Pr(z \in C)$, where C is the class of coiled coils [Ber95]. If $X = R_1, R_2, \ldots, R_l$ is a random subsequence selected from the universe of all known protein sequences, then

$$\begin{aligned}\Pr(z \in C) &= \Pr(X \in C | X = z) \\ &= \frac{\Pr(X = z | X \in C) \Pr(X \in C)}{\Pr(X = z)} \\ &\propto \frac{\Pr(R_1 = r_1 \wedge \ldots \wedge R_l = r_l | X \in C)}{\Pr(R_1 = r_1 \wedge \ldots \wedge R_l = r_l)}\end{aligned}$$

Using repeated applications of the definition of conditional probability, this is equal to:

$$\frac{\prod_{i=1}^{l-1} \Pr(R_i = r_i | R_{i+1} = r_{i+1} \wedge \ldots \wedge R_l = r_l \wedge X \in C) \cdot \Pr(R_l = r_l | X \in C)}{\prod_{i=1}^{l-1} \Pr(R_i = r_i | R_{i+1} = r_{i+1} \wedge \ldots \wedge R_l = r_l) \cdot \Pr(R_l = r_l)}. \quad (29.4)$$

To estimate these probabilities, it is necessary to make assumptions. For example, the simplest assumption is that the residues are independent of each other:

$$\Pr(R_i = r_i | R_{i+1} = r_{i+1} \wedge \ldots \wedge R_l = r_l \wedge X \in C) = \Pr(R_i = r_i | X \in C)$$

and

$$\Pr(R_i = r_i | R_{i+1} = r_{i+1} \wedge \ldots \wedge R_l = r_l) = \Pr(R_i = r_i).$$

Simplifying the previous equation with these assumptions gives

$$\Pr(z \in C) \propto \prod_{i=1}^{l} \frac{\Pr(R_i = r_i | X \in C)}{\Pr(R_i = r_i)},$$

and this is equivalent to the approach of [LvDS91].

In α-helices, a better assumption might be that a residue in position i is dependent on the next residue in the sequence $i+1$, as well as on those in positions $i+3$ and $i+4$, both of which are on the same face of the helix as position i (see Figure 29.5). This gives the following assumption: $\Pr(R_i = r_i | R_{i+1} = r_{i+1} \wedge \ldots \wedge R_l = r_l \wedge X \in C) = \Pr(R_i = r_i | R_{i+1} = r_{i+1} \wedge R_{i+3} = r_{i+3} \wedge R_{i+4} = r_{i+4} \wedge X \in C)$. However, this would require that $7^4 20^4$ parameters be estimated, and is not feasible in practice. The approach suggested in [Ber95] is to assume that $\Pr(R_i = r_i | R_{i+1} = r_{i+1} \wedge \ldots \wedge R_l = r_l \wedge X \in C)$ can be approximated by some function f (e.g., weighted average, minimum or maximum) over $\Pr(R_i = r_i | R_{i+1} = r_{i+1} \wedge X \in C)$, $\Pr(R_i = r_i | R_{i+3} = r_{i+3} \wedge X \in C)$ and $\Pr(R_i = r_i | R_{i+4} = r_{i+4} \wedge X \in C)$. More generally, if D is the set of dependencies (e.g., for helices, $D = \{1, 3, 4\}$ is the natural set of dependencies), then it is assumed that the probability of interest can be estimated as a function over the corresponding pairwise probabilities. In [BWW+95, BS97, SBK+98, SBK99], a geometric average over the pairwise probabilities is used.

The approach outlined above works well if the probabilities are estimated from a database representative of the types of coiled-coil structures that are to be predicted. However, the databases are heavily biased towards certain types of coiled coils. In [BS97], it is proposed that the basic method be used to iteratively scan a large database of sequences. Initially, the known database is used to estimate the required probabilities. Then, each sequence is scored using the framework described above, and this raw score is converted into a $(0, 1)$

probability p of its being a coiled coil. This probability is computed by fitting a Gaussian to the score distribution. In each iteration of the algorithm, a sequence is chosen with chance proportional to its probability of being coiled coil, and if chosen, its predicted coiled-coil residues will be used in the next iteration of the algorithm to update the probabilities. Single and pairwise frequencies are estimated in a Bayesian manner, with the initial estimates providing the prior. The iterative process continues until it stabilizes. This approach has been successful in identifying coiled-coil-like structures in histidine kinases [SBK+98] and viral membrane fusion proteins [SBK99], with crystal structures confirming several novel predictions [ZSMK00, MSK01].

Hidden Markov models (HMMs) have also been applied to coiled-coil recognition [BW95, DS02]. (For a general introduction to HMMs, see [DEKM00].) These approaches do not require that a fixed-length window be used, and thus may better predict shorter coiled-coil segments. Additionally, for coiled coils longer than a particular window length, HMMs can incorporate longer-range information than window-based approaches. In theory, HMMs permit modeling of interruptions in the heptad repeat pattern; however, in practice, such interruptions are severely penalized.

One HMM approach [DS02] builds a model of 64 states. There is a background state 0 corresponding to residues that do not take part in coiled coils. The other 63 states are denoted by a group number 1–9 and by a letter that refers to the heptad position. The first four groups model the first four residues in a coiled-coil segment, and the last four groups model the last four residues in a coiled-coil segment. The fifth group models internal coiled-coil residues. Each state corresponding to the same heptad repeat position is given the same emission probabilities. For groups 1–4 and 6–9, transition probabilities are specified to go from group i to group $i + 1$, with deviations from the heptad repeat pattern given some very small (though non-zero) chance. For group 0, self-transitions are allowed, as well as transitions to states in group 1. For state 5, there are transitions between states within this group as well as to states in group 6; in both cases, strong preference is given to transitions maintaining the heptad repeat. For any sequence, the prediction of whether each residue is in a coiled coil or not is given by the most likely state sequence through the HMM, given the sequence.

29.5.3 Predicting oligomerization

Natural coiled coils are known to exist as dimers, trimers, tetramers and pentamers. Attempts to predict oligomeric states of coiled-coil sequences have focused on differentiating between dimeric or trimeric coiled coils. In [WA95], amino acid frequencies at each heptad repeat position are computed for both dimeric and trimeric coiled coils, and normalized by the frequencies expected by chance. These give dimeric and trimeric propensities for each amino acid/heptad repeat pair. Each coiled-coil segment is then scored by summing the logs of the single frequency dimeric (and trimeric) propensities. Finally, the segment is predicted as dimeric if its dimeric propensity is higher than its trimeric one, and trimeric otherwise.

An alternate approach exploits pairwise residue correlations [WKB97] in predicting oligomerization state. This is a multidimensional scoring approach that uses the framework of [BWW+95]. Probabilities are estimated from a dimeric coiled-coil database, and then for $1 \leq d \leq 7$, each subsequence is scored assuming that dependencies exist between residues i and $i + d$. The analogous scores are computed using a trimeric database as well. Finally, a multidimensional score $\vec{s}$ for a subsequence z is converted to a probability of its being a

dimeric coiled coil by computing:

$$\frac{\Pr(\vec{s}|z \text{ is dimeric}) \cdot \Pr(\text{dimeric coiled coil})}{\Pr(\vec{s})}.$$

These probabilities are estimated by fitting multivariate Gaussians to the distributions of scores for dimeric coiled coils, trimeric coiled coils and non-coiled coils, and assuming a prior probability of dimeric, trimeric and non-coiled-coil residues. Trimer probabilities are computed similarly.

29.5.4 Structure-based predictions

The approaches outlined above have focused on statistical methods for predicting whether a given sequence takes part in a coiled-coil structure. There has also been work on predicting the high-resolution atomic structures of model coiled-coil systems using molecular mechanics. The earliest such attempts include [ZH93, VKBS94, DB94]. In [VKBS94], a hierarchical procedure is described to predict the structure of the GCN4 leucine zipper; a backbone root-mean-squared deviation (RMSD) of 0.81 Å is obtained when predicting the dimeric GCN4 leucine zipper. In [DB94], dimeric and tetrameric variants of GCN4 are considered, and an RMSD of 0.73 Å is obtained for residues in the dimerization interface.

The coiled-coil backbone can be parameterized [Cri53], and [HTK95] show how to exploit this parameterization in order to incorporate backbone flexibility in predicting structures. Coiled-coil backbones can be described by specifying the superhelical radius R_0, the superhelix frequency ω_0, the α-helical radius R_1, the helical frequency ω_1, and the rise per amino acid in the α-helix d. The heptad repeat fixes ω_1 to be $4\pi/7$ radians per amino acid, so that seven residues complete two full turns relative to the superhelical axis, and place every seventh residue in the same local environment. Additionally, it may be assumed that the helices making up the coiled-coil are regular and symmetric, and so d can be fixed to be the rise per amino acid of a regular α-helix (1.52 Å) and R_1 can be fixed to be the C_α radius of a regular α-helix (2.26 Å). The remaining parameters can then be varied to enumerate backbone conformations. Side chains are then positioned on these backbones via energy minimization. This approach has resulted in predictions with root-mean-square deviation from crystal structures of less than 0.6 Å when considering hydrophobic **a** and **d** position residues for three GCN4 variants (2-stranded, 3-stranded and 4-stranded). Additionally, a novel coiled-coil backbone consisting of a *right-handed* superhelical twist and an 11-mer repeat has been designed using the parameterized-backbone approach [HPT$^+$98]. A parameterized-backbone approach has also been used to predict the hydrophobic dimerization interface of six designed heterodimeric coiled coils [KMTK01], as well as to predict the differences in stabilities of these constructs. In this approach, for each backbone, all near-optimal packings of side chains are identified, and these structures are then relaxed via energy minimization [BBO$^+$83] to find the minimum energy backbone and side-chain conformations.

29.5.5 Predicting coiled-coil protein interactions

As outlined above, effective sequence-based prediction methods exist for recognizing single helices that take part in coiled coils. Since coiled coils are made up of two or more helices that interact with each other, a natural next step in predicting their structures is to try to predict which helices are interacting with each other. Since these helices may be in different protein sequences, this begins to address the problem of predicting protein-protein

interactions. This is an important problem as protein-protein interactions play a central role in many cellular functions. Furthermore, the difficulty of computationally predicting protein structures suggests a strategy of concentrating first on interactions mediated by specific interfaces of known geometry.

Early approaches towards predicting coiled-coil interaction specificity have counted the number of favorable and unfavorable electrostatic interactions to make some specific predictions about the nature of particular coiled-coil protein-protein interactions [Par77, MS75, VHB93]; however, it is known that many other factors play a role in coiled-coil specificity (e.g., [ORK92, LK95, HZKA93]) and thus such simple approaches are limited in their applicability.

An alternative approach represents coiled coils in terms of their interhelical residue interactions and derives a "weight" that indicates how favorable each residue-residue interaction is [SK01, FKS04]. Unlike the other sequence-based approaches outlined in this chapter, this approach uses not only sequence and structural data, but also experimental data. This use of experimental data is critical to its performance. The approach has thus far been applied only to predicting partners for helices taking part in dimeric coiled coils. In dimeric coiled coils, residues at the **a**, **d**, **e**, and **g** positions form the protein-protein interface [OKKA91, GH95] (see Figure 29.5). Experimental studies show that specificity is largely driven by interactions between residues at these core positions (e.g., see [VMA$^+$02]). The method further assumes that considering interhelical interactions among these residues in a pairwise manner is sufficient.[7] Based on structural features of the interhelical interface [OKKA91, GH95] as well as experiments on determinants of specificity (e.g., [ORK92, LK95, VHB93]), the following seven interhelical interactions are assumed to govern partnering in coiled coils:

$$\mathbf{a}_i\mathbf{d}'_i,\ \mathbf{d}_i\mathbf{a}'_{i+1},\ \mathbf{d}_i\mathbf{e}'_i,\ \mathbf{g}_i\mathbf{a}'_{i+1},\ \mathbf{g}_i\mathbf{e}'_{i+1},\ \mathbf{a}_i\mathbf{a}'_i,\ \mathbf{d}_i\mathbf{d}'_i. \tag{29.5}$$

The prime differentiates the two strands and the subscript denotes the relative heptad number (e.g., the first interaction, $\mathbf{a}_i\mathbf{d}'_i$, is between the **a** position in the i-th heptad of one helix and the **d** position in the same heptad of the other helix).

Consequently, each coiled-coil structure is represented as a 2800-dimensional vector $\vec{x}$, the entries of which tabulate the occurrences of amino-acid pairs in the above interactions. Specifically, entry $x_{(p,q),i,j}$ indicates the number of times amino acids i and j appear across the helical interface in positions p and q, respectively.

Scoring framework. For each possible interhelical interaction, the method needs a weight $w_{(p,q),i,j}$ that denotes how favorable the interaction is between amino acid i in position p and amino acid j in position q. A potential coiled coil represented by $\vec{x}$ is then scored by computing $\vec{w} \cdot \vec{x}$ where $\vec{w}$ is a vector of such weights. Initially this weight vector $\vec{w}$ is not known; however, these weights should satisfy certain constraints.

Experimental information on relative coiled-coil stability (e.g, the observation that coiled coil $\vec{x}$ is more stable than coiled coil $\vec{y}$) is used to constrain the weight vector $\vec{w}$ by requiring that

$$\vec{w} \cdot \vec{x} \quad > \quad \vec{w} \cdot \vec{y}. \tag{29.6}$$

Additionally, sequences known to form coiled coils should score higher than those that do

[7] It is possible to consider three or more amino acids at a time but this would require a larger coiled-coil database.

not:

$$\vec{w} \cdot \vec{x} \;>\; 0, \text{ for all coiled coils } \vec{x}, \tag{29.7}$$

$$\vec{w} \cdot \vec{y} \;<\; 0, \text{ for all non-coiled coils } \vec{y}. \tag{29.8}$$

These constraints are similar to those seen most often in machine learning settings.

Finally, knowledge about specific weight elements can be directly incorporated. For example, say it is favorable to have a lysine in a **g** position in one helix with a glutamic acid in the following position **e** in the other helix, but not favorable to have glutamic acid in both these positions (i.e., **g**-**e** K E is "better than" **g**-**e** E E). Then the following should be true:

$$w_{(g,e),K,E} \;>\; 0, \quad w_{(g,e),E,E} \;<\; 0. \tag{29.9}$$

Indexing each constraint with i, the above constraints (equations 29.6–29.9) can be rewritten using vectors $\vec{z}^{(i)}$, such that $\vec{w}$ is constrained to satisfy $\vec{w} \cdot \vec{z}^{(i)} > 0$. Including non-negative slack variables ϵ_i to allow for errors in sequence or experimental data, each constraint can then be relaxed as $\vec{w} \cdot \vec{z}^{(i)} \geq -\epsilon_i$. The goal is to find $\vec{w}$ and $\vec{\epsilon}$ such that each constraint is satisfied and $\sum \epsilon_i$ is minimized. Trade-offs between training and generalization error suggest the approach of support vector machines (SVMs) [Vap98, Bur98], in which the following quadratic objective function (for some constant C) is minimized, subject to a variation of the previously described set of linear constraints:

$$\frac{1}{2} \parallel \vec{w} \parallel^2 + C(\sum \epsilon_i)$$

subject to

$$\begin{aligned} \vec{w} \cdot \vec{z}^{(i)} &\geq 1 - \epsilon_i \quad \forall i \\ \epsilon_i &\geq 0 \qquad\quad \forall i \end{aligned}$$

Differences between this approach and the traditional application of SVMs include constraints on specific elements of the weight vector, and constraints about the relative "score" of different interactions.

This approach has been tested on a near-complete set of coiled-coil interactions among human and yeast leucine zipper bZIP transcription factors [NK03], and identifies 70% of strong interactions while maintaining that 92% of predictions are correct [FKS04]. Though genomic approaches to predicting protein partners have had some success (e.g., [DSHB98, MPN+99, EIKO99, GBJ+00, RM03, JYG+03, YLL+04]), as have structure-based threading methods [AR02, LLS02], the coiled coil is the first interaction interface for which these types of high-confidence, large-scale computational predictions can be made.

29.5.6 Promising future directions

Since secondary structure prediction methods improved considerably by incorporating evolutionary information, the next obvious step in improving recognition of helices taking part in coiled coil structures is to use homologous sequences. For predicting coiled-coil interactions, however, homologous sequences can show very different interaction specificity [NK03], and thus it is not obvious how to exploit evolutionary information in this context. Additionally, while methods have been developed for predicting whether a coiled coil helix is likely to take part in either a dimeric and trimeric structure [WA95, WKB97], there are no methods for predicting higher-order oligomerization states or for predicting whether the helices interact in a parallel or anti-parallel manner. Finally, methods for predicting coiled-coil protein interactions have focused on parallel, 2-stranded coiled coils, and novel approaches are needed for predicting coiled-coil protein interactions more generally.

29.6 Conclusions

In this chapter, we have reviewed the basic computational methods used to predict protein secondary structure, as well as β hairpin and coiled coil supersecondary structures. Of these problems, secondary structure prediction has been the most widely studied, and almost all successful methods for predicting tertiary structure rely on predictions of secondary structure (e.g.,see [ASHR03]). As methods for predicting other types of local structure improve, they are likely to play an increasing role in tertiary structure prediction methods. More recently, effective methods for predicting other types of β-structures, including β-helices [BCM+01]and β-trefoils [MSK+04], have also been developed, and these types of specialized computational approaches provide a new means for predicting protein tertiary structure. Finally, protein interactions are also mediated by various well-characterized structural motifs (e.g., see [PRN02]), and as demonstrated with the coiled coil, a promising approach for making high-confidence predictions of protein interactions and quartenary structure is to focus first on interactions mediated by specific, local structural interfaces.

Acknowledgments

The author thanks Carl Kingsford, Elena Nabieva and Elena Zaslavsky for helpful discussions, and the NSF for PECASE award MCB–0093399.

References

[AHSW61] C. Anfinsen, E. Haber, M. Sela, and F. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences (USA)*, 47:1309+, 1961.

[AMS+97] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389+, 1997.

[Anf73] C. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223+, 1973.

[AR02] P. Aloy and R. Russell. Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences*, 99:5896+, 2002.

[ASHR03] P. Aloy, A. Stark, C. Hadley, and R. Russell. Prediction without templates: New folds, secondary structure, and contacts in CASP5. *Proteins: Structure, Function and Bioinformatics*, 53:436+, 2003.

[BBF+99] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri. Exploiting the past and present in secondary structure prediction. *Bioinformatics*, 15:937+, 1999.

[BBO+83] B. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.

[BBRV98] E. Bornberg-Bauer, E. Rivals, and M. Vingron. Computational approaches to identify leucine zippers. *Nucleic Acids Research*, 26:2740+, 1998.

[BCM+01] P. Bradley, L. Cowen, M. Menke, J. King, and B. Berger. BETAWRAP:

Successful prediction of parallel β-helices from primary sequence reveals an association with many microbial pathogens. *Proceedings of the National Academy of Sciences*, 98:14819+, 2001.

[BCM+03] P. Bradley, D. Chivian, J. Meiler, K. Misura, C. Rohl, W. Schief, W. Wedemeyer, O. Schueler-Furman, P. Murphy, J. Schonbrun, C. Strauss, and D. Baker. Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation. *Proteins: Structure, Function and Genetics*, 53:457+, 2003.

[Ber95] B. Berger. Algorithms for protein structural motif recognition. *Journal of Computational Biology*, 2:125+, 1995.

[BK89] V. Brendel and S. Karlin. Too many leucine zippers? *Nature*, 341:574+, 1989.

[BL93] S. H. Bryant and C. E. Lawrence. An empirical energy function for threading protein sequence through the folding motif. *Proteins: Structure, Function and Genetics*, 16:92–112, 1993.

[BLE91] J. Bowie, R. Luthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-diensional structure. *Science*, 253:164+, 1991.

[BS97] B. Berger and M. Singh. An iterative method for improved protein structural motif recognition. *Journal of Computational Biology*, 4(3):261+, 1997.

[BT99] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, Inc., 1999.

[Bur98] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121+, 1998.

[BW95] B. Berger and D. Wilson. Improved algorithms for protein motif recognition. In *Symposium on Discrete Algorithms*, pages 58+. SIAM, January 1995.

[BWF+00] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhatand H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235+, 2000.

[BWW+95] B. Berger, D. B. Wilson, E. Wolf, T. Tonchev, M. Milla, and P. S. Kim. Predicting coiled coils using pairwise residue correlations. *Proceedings of the National Academy of Sciences*, 92:8259+, 1995.

[CB97] K.-C. Chou and J. Blinn. Classification and prediction of β-turn types. *Journal of Protein Chemistry*, 16:575+, 1997.

[CB99a] J. Cuff and G. Barton. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Structure, Function and Genetics*, 40:502+, 1999.

[CB99b] J. Cuff and G. Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Structure, Function and Genetics*, 34:508+, 1999.

[CC99] Y.-D. Cai and K.-C. Chou. Artificial neural network model for predicting α-turn types. *Analytical Biochemistry*, 268:407+, 1999.

[CF74] P. Chou and G. Fasman. Prediction of protein conformation. *Biopolymers*, 13:211+, 1974.

[CF79] P. Chou and G. Fasman. Prediction of β-turns. *Biophysical Journal*, 26:367+, 1979.

[CFLC03] Y.-D. Cai, K.-Y. Feng, Y.-X. Li, and K. C. Chou. Support vector machine for predicting α-turn types. *Peptides*, 24:629+, 2003.

[Cho97] K.-C. Chou. Prediction of β-turns. *Journal of Peptide Research*, 49:120+, 1997.

[Cho00] K.-C. Chou. Prediction of tight turns and their types in proteins. *Analytical*

Biochemistry, 286:1+, 2000.

[CLL+03] Y.-D. Cai, X.-J. Liu, Y.-X. Li, X.-B. Xu, and K.-C. Chou. Prediction of β turns with learning machines. *Peptides*, 24:665+, 2003.

[Coh98] C. Cohen. Why fibrous proteins are romantic. *Journal of Structural Biology*, 112:3+, 1998.

[Cri53] F. H. C. Crick. The packing of α-helices: simple coiled coils. *Acta Crystallographica*, 6:689, 1953.

[Cyb89] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303+, 1989.

[DB94] W. DeLano and A. Brunger. Helix packing in proteins: prediction and energetic analysis of dimeric, trimeric, and tetrameric GCN4 coiled coil structures. *Proteins: Structure, Function and Genetics*, 20:105+, 1994.

[DEKM00] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 2000.

[dlCHST02] X. de la Cruz, E. Hutchinson, A. Shepherd, and J. Thornton. Toward predicting protein topology: an approach to identifying β hairpins. *Proceedings of the National Academy of Sciences*, 99:11157+, 2002.

[DS02] M. Delorenzi and T. Speed. An HMM model for coiled-coil domains and a comparison with pssm-based predictions. *Bioinformatics*, 18:617+, 2002.

[DSHB98] T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9):324+, 1998.

[EIKO99] A.J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86+, 1999.

[EMRP+01] V. Eyrich, M. Marti-Renom, D. Przybylski, M. Madhusudhan, A. Fiser, F. Pazos, A. Valencia, A. Sali, and B. Rost. EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, 17:1242+, 2001.

[EWT84] D. Eisenberg, R. Weiss, and T. Terwilliger. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proceedings of the National Academy of Sciences (USA)*, 81:140+, 1984.

[FA95] D. Frishman and P. Argos. Knowledge-based secondary structure assignment. *Proteins: Structure, Function and Genetics*, 23:566+, 1995.

[FA96] D. Frishman and P. Argos. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Engineering*, 9:133+, 1996.

[FA97] D. Frishman and P. Argos. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins: Structure, Function and Genetics*, 27:329+, 1997.

[Fan61] R. Fano. *Transmission of Information*. Wiley, New York, 1961.

[FKS04] J. Fong, A. E. Keating, and M. Singh. Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biology*, 5(2):R11, 2004.

[GBJ+00] C. Goh, A. Bogan, M. Joachimiak, D. Walther, and F. Cohen. Co-evolution of proteins with their interaction partners. *J. Mol. Biol*, 299:283+, 2000.

[GGR87] J. Gibrat, J. Garnier, and B. Robson. Further developments of protein secondary structure prediction using information theory: new parameters and consideration of residue pairs. *Journal of Molecular Biology*, 198:425+, 1987.

[GH95] J. Glover and S. Harrison. Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature*, 373:257+, 1995.

[GOR78] J. Garnier, D. Osguthorpe, and B. Robson. Analysis and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, 120:97+, 1978.

[HK89] L. H. Holley and M. Karplus. Protein secondary structure prediction with a neural net. *Proceedings of the National Academy of Sciences (USA)*, 86:152+, 1989.

[HPT+98] P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, and P. S. Kim. High-resolution protein design with backbone freedom. *Science*, 282:1462+, 1998.

[HS01] S. Hua and Z. Sun. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *Journal of Molecular Biology*, 308:397+, 2001.

[HT94] E. G. Hutchinson and J. Thornton. A revised set of potentials for β-turn formation in proteins. *Protein Science*, 3:2207+, 1994.

[HTK95] P. B. Harbury, B. Tidor, and P. S. Kim. Predicting protein cores with backbone freedom: Structure prediction for coiled coils. *Proceedings of the National Academy of Sciences*, 92:8408+, 1995.

[Hur95] H. Hurst. Transcription factors 1: bZIP proteins. *Protein Profile*, 2(2):101+, 1995.

[HVSB96] J. Hirst, M. Vieth, J. Skolnick, and C. Brooks. Predicting leucine zipper structures from sequence. *Protein Engineering*, 9:657+, 1996.

[HZKA93] P. B. Harbury, T. Zhang, P. S. Kim, and T. Alber. A switch between two-, three- and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science*, 262:1401+, November 1993.

[Jon99] D. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292:195+, 1999.

[JTT92] D. Jones, W. Taylor, and J. Thornton. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.

[JYG+03] R. H. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. Krogan, S. Chung, A. Emili, M. Snyder, J. Greenblatt, and M.Gerstein. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302:449+, 2003.

[KBH98] K. Karplus, C. Barret, and R. Hughey. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14:846+, 1998.

[Kim04] S. Kim. Protein β-turn prediction using nearest-neighbor method. *Bioinformatics*, 20:40+, 2004.

[KKD+03] K. Karplus, R. Karchin, J. Draper, J. Casper, Y. Mandel-Gutfreund, M. Diekhans, and R. Hughey. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins: Structure, Function and Genetics*, 53:491+, 2003.

[KMB04] M. Kuhn, J. Meiler, and D. Baker. Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. *Proteins: Structure, Function and Bioinformatics*, 54:282+, 2004.

[KMTK01] A. E. Keating, V. Malashkevich, B. Tidor, and P. S. Kim. Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. *Proceedings of the National Academy of Sciences*, 98(26):14825+, 2001.

[KOS+00] R. King, M. Ouali, A. Strong, A. Aly, A. Elmaghraby, M. Kantardzic, and D. Page. Is it better to combine predictions? *Protein Engineering*, 13:15+, 2000.

[KR02] H. Kaur and G. Raghava. An evaluation of β-turn prediction methods. *Bioin-*

formatics, 18:1508+, 2002.

[KR03a] H. Kaur and G. Raghava. A neural-network based method for prediction of γ-turns in proteins from multiple sequence aligments. *Protein Science*, 12:923+, 2003.

[KR03b] H. Kaur and G. Raghava. Prediction of β-turns in proteins from multiple aligment using neural network. *Protein Science*, 12:627+, 2003.

[KR04] H. Kaur and G. Raghava. A neural network method for prediction of β-turn types in proteins using evolutionary information. *Bioinformatics*, 20:2751+, 2004.

[KS83a] W. Kabsch and C. Sander. A dictionary of protein secondary structure. *Biopolymers*, 22:2577+, 1983.

[KS83b] W. Kabsch and C. Sander. How good are predictions of protein secondary structure? *FEBS Lett.*, 155:179+, 1983.

[KS96] R. King and M. Sternberg. Identification and application of the concepts important for accuracte and reliable protein secondary structure prediction. *Protein Science*, 5:2298+, 1996.

[KTJG02] A. Kloczkowski, K.-L. Ting, R. Jernigan, and J. Garnier. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Journal of Molecular Biology*, 49:154+, 2002.

[KWK+03] L. Kinch, J. Wrabl, S. Krishna, I. Majmudar, R. Sadreyev, Y. Qi, J. Pei, H. Cheng, and N. Grishin. CASP5 assessment of fold recognition target predictions. *Proteins: Structure, Function and Bioinformatics*, 53:395+, 2003.

[Les01] A. Lesk. *Introduction to protein architecture*. Oxford University Press, 2001.

[LK95] K. Lumb and P. S. Kim. A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry*, 34:8642+, 1995.

[LL93] T. Li and E. Lander. Protein secondary structure prediction using nearest-neighbor methods. *Journal of Molecular Biology*, 232:1117+, 1993.

[LLS02] L. Lu, H. Lu, and J. Skolnick. Multiprospector: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, 49(3):1895+, 2002.

[LMS71] P. Lewis, F. Momany, and H. Scheraga. Folding of polypeptide chains in proteins: a proposed mechanism for folding. *Proceedings of the National Academy of Sciences*, 68:2293+, 1971.

[LMS73] P. Lewis, F. Momany, and H. Scheraga. Chain reversals in proteins. *Biochimica et Biophysica Acta*, 303:211+, 1973.

[LRG86] J. Levin, B. Robson, and J. Garnier. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Letters*, 205(2):303+, 1986.

[LvDS91] A. Lupas, M. van Dyke, and J. Stock. Predicting coiled coils from protein sequences. *Science*, 252:1162+, 1991.

[Mat75] B. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405:442+, 1975.

[MFS89] M. McGregor, T. Flores, and M. Sternberg. Prediction of β-turns in proteins using neural networks. *Protein Engineering*, 2:521+, 1989.

[MFZH03] J. Moult, K. Fidelis, A. Zemla, and T. Hubbard. Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins: Structure, Function, and Genetics*, 53:334+, 2003.

[MK96] D. Minor and P. S. Kim. Context-dependent secondary structure formation of a designed protein sequence. *Nature*, 380(6576):730+, 1996.

[MPN+99] E. Marcotte, M. Pellegrini, H. Ng, D. Rice, T. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285:751+, 1999.

[MS75] A. McLachlan and M. Stewart. Tropomyosin coiled-coil interactions: Evidence for an unstaggered structure. *Journal of Molecular Biology*, 98:293+, 1975.

[MSK01] V. Malashkevich, M. Singh, and P. S. Kim. The trimer-of-hairpins motif in viral membrane-fusion proteins: Visna virus. *Proceedings of the National Academy of Sciences*, 98:8502+, 2001.

[MSK+04] M. Menke, E. Scanlon, J. King, B. Berger, and L. Cowen. Wrap-and-pack: A new paradigm for beta structural motif recognition with application to recognizing beta trefoils. In *Proceedings of the 8th Annual International Conference on Computational Molecular Biology*, pages 298+. ACM, 2004.

[NK03] J. R. S. Newman and A. E. Keating. Comprehensive identification of human bZIP interactions using coiled-coil arrays. *Science*, 300:2097+, 2003.

[NO86] K. Nishikawa and T. Ooi. Amino acid sequence homology applied to prediction of protein secondary structure, and joint prediction with existing methods. *Biochimica et Biophysica Acta*, 871(1):45+, 1986.

[OKKA91] E. O'Shea, J. Klemm, P. S. Kim, and T. Alber. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science*, 254:539+, October 1991.

[ORK92] E. O'Shea, R. Rutkowski, and P. S. Kim. Mechanism of specificity in the fos-jun oncoprotein heterodimer. *Cell*, 68:699+, 1992.

[Par77] D. A. D. Parry. Sequences of α-keratin: Structural implication of the amino acid sequences of the type I and type II chain segments. *Journal of Molecular Biology*, 113:449+, 1977.

[Par82] D. A. D. Parry. Coiled coils in alpha-helix-containing proteins: analysis of residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins. *Bioscience Reports*, 2:1017+, 1982.

[PPRB02] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks. *Proteins: Structure, Function and Genetics*, 47:228+, 2002.

[PRN02] T. Pawson, M. Raina, and P. Nash. Interaction domains: from simple binding events to complex cellular behavior. *FEBS Letters*, pages 2+, 2002.

[QS88] N. Qian and T. Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology*, 202(4):865+, 1988.

[RHW86a] D. Rumelhart, G. Hinton, and R. Williams. Learning internal representations by error propagation. In D. Rumelhart and J. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 323, pages 318–362. MIT Press, Cambridge, MA, 1986.

[RHW86b] D. Rumelhart, G. Hinton, and R. Williams. Learning representations by back-propagating errors. *Nature*, 323:533+, 1986.

[Ric81] J. Richardson. The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry*, 34:167+, 1981.

[RM03] A. Ramani and E. Marcotte. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *Journal of Molecular Biology*, 327:273+, 2003.

[Ros01] B. Rost. Protein secondary structure prediction continues to rise. *Journal of Structural Biology*, 134:204+, 2001.

[RS93] B. Rost and C. Sander. Prediction of protein secondary structure at better

than 70%. *Journal of Molecular Biology*, 232:584+, 1993.

[RSS94a] B. Rost, C. Sander, and R. Schneider. PhD: an automatic mail server for protein secondary structure prediction. *Computer Applications in Biosciences*, 10:53+, 1994.

[RSS94b] B. Rost, C. Sander, and R. Schneider. Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology*, 235:13+, 1994.

[SBK+98] M. Singh, B. Berger, P. S. Kim, J. Berger, and A. Cochran. Computational learning reveals coiled coil-like motifs in histidine kinase linker domains. *Proceedings of the National Academy of Sciences*, 95:2738+, March 1998.

[SBK99] M. Singh, B. Berger, and P. S. Kim. Learncoil-VMF: Computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins. *Journal of Molecular Biology*, 290:1031+, 1999.

[SGT99] A. Shepherd, D. Gorse, and J. Thornton. Prediction of the location and type of β-turns in proteins using neural networks. *Protein Science*, 8:1045+, 1999.

[Sip90] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force. *Journal of Molecular Biology*, 213:859–883, 1990.

[SK01] M. Singh and P. S. Kim. Towards predicting coiled-coil protein interactions. In *Proceedings of the 5th Annual International Conference on Computational Molecular Biology*, pages 279+. ACM, 2001.

[SKHB97] K. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, 268:209+, 1997.

[SS95] A. Salamov and V. Solovyev. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignment. *Journal of Molecular Biology*, 247:11+, 1995.

[ST85] B. Sibanda and J. Thornton. Beta-hairpin families in globular proteins. *Nature*, 316:170+, 1985.

[TM03] A. Tramontano and V. Morea. Assessment of homology-based predictions in CASP5. *Proteins: Structure, Function, and Genetics*, 53:352+, 2003.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[Ven68] C. Venkatachalam. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers*, 6:1425+, 1968.

[VHB93] C. Vinson, T. Hai, and S. Boyd. Dimerization specificy of the leucine zipper-containing bZIP motif on DNA binding: prediction and rational design. *Genes and Development*, 7(6):1047+, 1993.

[VKBS94] M. Vieth, A. Kolinski, C. L. Brooks, and J. Skolnick. Prediction of the folding pathways and structure of the GCN4 leucine zipper. *Journal of Molecular Biology*, 237:361+, 1994.

[VMA+02] C. Vinson, M. Myakishev, A. Acharya, A. Mir, J. R. Moll, and M. Bonovich. Classification of human bZIP proteins based on dimerization properties. *Molecular and Cellular Biology*, 22(18):6321–6335, 2002.

[WA95] D. Woolfson and T. Alber. Predicting oligomerization states of coiled coils. *Protein Science*, 4:1596–1607, 1995.

[WKB97] E. Wolf, P. S. Kim, and B. Berger. Multicoil: A program for predicting two- and three-stranded coiled coils. *Protein Sci.*, 6:1179+, 1997.

[WT88] C. Wilmot and J. Thornton. Analysis and prediction of the different types of β-turn in proteins. *Journal of Molecular Biology*, 203:221+, 1988.

[WT90] C. Wilmot and J. Thornton. β-turns and their distortions: a proposed new

nomenclature. *Protein Engineering*, 3:479+, 1990.

[WW01] J. Walshaw and D. Woolfson. Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *Journal of Molecular Biology*, 307:1427+, 2001.

[YLL+04] H. Yu, N. Luscombe, H. Lu, X. Zhu, Y. Xia, J. Han, N. Bertin, S. Chung, M. Vidal, and M. Gerstein. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Research*, 14:1107+, 2004.

[ZBTS87] M. Zvelebil, G. Barton, W. Taylor, and M. Sternberg. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *Journal of Molecular Biology*, 195(4):957+, 1987.

[ZC97] C.-T. Zhang and K. C. Chou. Prediction of β-turns in proteins by 1–4 and 2–3 correlation model. *Biopolymers*, 41:673+, 1997.

[ZH93] L. Zhang and J. Hermans. Molecular dynamics study of structure and stability of a model coiled coil. *Proteins: Structure, Function and Genetics*, 16:384+, 1993.

[ZSMK00] X. Zhao, M. Singh, V. Malashkevich, and P. S. Kim. Structural characterization of the human respiratory syncytial virus fusion protein core. *Proceedings of the National Academy of Sciences*, 97:14172+, 2000.

[ZVFR99] A. Zemla, C. Venclovas, K. Fidelis, and B. Rost. A modified definition of Sov, a segment-based measure for protein structure prediction assessment. *Proteins: Structure, Function and Genetics*, 34:220+, 1999.