Introduction

Why Predict Protein Structure?

Globular proteins owe their importance to their unique **tertiary structure**. This allows them to bind, be regulated by and transport smaller molecules; interact with and regulate larger ones; and catalyze biochemical reactions. Structural biochemists who use x-ray crystallography, nuclear magnetic resonance, circular dichroism, and other physical techniques to predict tertiary structure do so through measurements on folded protein molecules. However, the primary sequence of a potential protein can now be determined from DNA and many such sequences are being reported. Which of those are most likely to warrant further study can often be determined with knowledge of the tertiary structure. Thus it would be extremely useful if tertiary structure could be predicted directly from **primary structure**.

Each tertiary structure is determined by a primary sequence, but exactly how (**the protein folding problem**) is the subject of much current research. Tertiary structure can be considered an aggregate of alpha helices, beta strands and turns, elements of **secondary structure**. If secondary structure could be accurately predicted from primary sequence, one would "only" need to correctly pack secondary structural elements - a seemingly less complicated task- and the protein folding problem would be solved (for an introduction to the packing problem, see Nagano, 1989). This is one approach to protein structure prediction from primary sequence.

Alternative approaches use empirical techniques or molecular mechanics and dynamics to predict tertiary structure without necessarily first predicting secondary structure. This chapter summarizes progress in protein structure prediction, emphasizing methods for predictions in globular proteins since Fasman (1989a). Other reviews include Nishikawa and Noguchi (1991); Garnier and Levin (1991); Swindells and Thornton (1991); Benner (1992, 1993); Rost *et al.* (1993); and Cohen and Cohen (1994). Structure prediction for membrane-bound proteins is reviewed by Heijne (1994). The chapter includes a list of Resources for protein structure prediction, and a Glossary. Words and phrases included in the glossary are printed in **boldface type** when they first appear in the text.

Background and Overview

Prediction of structure began in the 1960s when the first protein crystal structures were available for detailed study. Prothero (1966) reported certain amino acids could be used to predict helices in myoglobin and hemoglobin. Schiffer & Edmunson (1967) developed the **helical wheel** both to predict helical potential and, if a helix is present, to indicate the presence of a hydrophobic region. Ptitsyn (1969) studied secondary structures of seven globular proteins and found certain amino acids were partitioned differently between helical and non-helical sections; his conclusions agreed with Prothero. More details of the early history in this field is given in Fasman (1989b).

The most commonly used secondary structure prediction methods today were developed in the following decade. These include the **statistical methods** of Chou & Fasman (1974a, 1974b) and of Garnier and co-workers (Garnier *et al.*, 1978; Garnier and Robson, 1989). Alternative approaches also applied to the problem of predicting secondary and tertiary structure include **neural networks** (for example, Qian & Sejnowski, 1988) and **molecular modeling** (for example, Bruccoleri and Karplus, 1987).

Any method of structure prediction must be first tested on sequences with known tertiary structure. Early studies were limited by the small number of proteins with known structure, and of necessity used all that were available. As more structures were reported, researchers had the choice of continuing to base their methods on all known globular proteins, and in that sense seek universal rules, or to limit their universe to certain **structural classes** or even single protein families. Today both approaches are being actively explored. Predictive studies limited to a single protein family are called **homology modeling**.

Evaluating the Accuracy of Predictions

To compare various methods of protein structure prediction requires a gold standard of structure. Structures derived from X-ray or nmr measurements are usually the standards for tertiary structure prediction. Many published tertiary structures include secondary structure assignments, but these can be incomplete and subjective. Thus when secondary structure is predicted, it is often preferable to reassign secondary structure from tertiary coordinates using a computer program such as DSSP (Kabsch and Sander, 1983 TARGET = "refs").

Accuracy of tertiary structure predictions are usually measured by comparing the coordinates for correct and predicted structures using the **root mean square (r.m.s.) deviation**. Let x_i stand for a set of atomic coordinates for one atom in a (possibly known) structure, and y_i for the corresponding atom in a second (possibly predicted) structure. One can mathematically transform the set of y_i coordinates to Y_i such that the sum of the squares of the distance deviations

$$\sum |x_i - Y_i|^2 \quad (1)$$

is a minimum. Then the r.m.s. deviation is defined as:

$$\Delta R = \sqrt{\frac{\sum |x_i - Y_i|^2}{N}} \quad (2)$$

where N is the total number of atoms in the structure (for further discussion of r.m.s., see Lesk (1991)). Cohen and Kuntz (1989) emphasize that such r.m.s. measurements must be compared to measurements on random structures constrained to pack in a sphere. Cohen and Sternberg (1980) developed an equation for determining such random r.m.s. deviations:

$$\Delta R = 0.0468N + 9.25 \quad (3)$$

where N here is the number of residues in the protein sequence. When such a comparison is made, early predicted tertiary structures are little better than random.

Another method to analyze the spacial errors between two tertiary protein structures is to use **volume overlap integrals** (Schiffer *et al.*, 1990). The two structures are superimposed by overlapping their C_α backbones. The volume of a particular residue is calculated by extending the atomic coordinates of each atom into a sphere with a radius equal to its van der Waals radius. The percentage volume overlap between two residues is determined by the volume overlap between the predicted residue and the residue in the crystal structure.

The accuracy of sequence patterns developed in certain types of homology modeling is best measured by developing two sets of test sequences, one of the sequences which are contain the structural feature under study (knowns) and one of representative sequences which do not contain it (controls). Then pattern accuracy can be assessed by counting the number of correct matches (true positives, TP) where it is found in the knowns; correct non-matches (true negatives, TN) where it is not found in the controls; incorrect matches (false positives, FP) where it is found in the controls; and incorrect non-matches (false negatives, FN) where it is not found in the knowns. Two measures of pattern accuracy are **sensitivity** = TP/(TP + FN), and **specificity** = TN/(TN + FP), and both must be calculated for a pattern to be evaluated. The validity of sequence patterns is further discussed by Lathrop *et al.* (1993).

Even given a standard for secondary structure, the best measure of accuracy for secondary structure predictions is not as clear. A performance measure that accounts for both over- and under-prediction is the **Mathews correlation coefficient** developed by Mathews (1975). For the structure type *a*, the correlation coefficient is defined by

$$C_a = \frac{(p_a n_a) - (u_a o_a)}{\sqrt{(n_a + u_a)(n_a + o_a)(p_a + u_a)(p_a + o_a)}} \quad (4)$$

where p_a is the number of correctly predicted cases, n_a is the number of correctly rejected cases, o_a is the number of overpredicted cases, and u_a is the number of underpredicted cases. A more frequently used measure, **single residue accuracy**, is the number of residues correctly predicted to contain a structure divided by the number of residues that do contain that structure. To determine overall accuracy, this can be summed over the number of different structures or states predicted, usually either three (helix, strand, other) or four (helix, strand, turn, coil).

Since turns and surface (random coil) loops are frequently interchanged in homologous proteins, three-state accuracy is arguably the better measure. Also four-state values can easily be converted into three-state ones. **Three-state single residue accuracy (Q_3)** is:

$$Q_3 = \frac{P_\alpha + P_\beta + P_{coil}}{N} \quad (5)$$

where *N* is the total number of predicted residues and P_a is the number of correctly predicted secondary structures of type *a*. Q_3 values of from 0.5 to 0.7 (50-70% accuracy) have been reported for Chou-Fasman; Garnier, Osguthorpe and Robson (GOR); and other current methods.

Jenny and Benner (1994a) list several deficiencies of the Q_3 score for the evaluation of secondary structure predictions and recommend several scores to be reported in addition: the three individual single residue scores which make up Q_3; a score that reflects the number of serious errors (those where helix is mistaken for strand and vice versa); and a score for accuracy of prediction for each individual structural element. They also recommend three guidelines for comparing predictions. First, compare a consensus prediction (one made from several homologous sequences) with a "consensus" experimental structure, or, where multiple experimental structures do not exist, lower the "target score" for a perfect prediction to reflect the diversity found in secondary structure in homologous proteins. Second, compare prediction methods only with others that are similar (for example, all completely sequence-based, or all incorporating the same type of experimental data). Third, compare predictions made before a structure is known (*de novo* predictions) only other such predictions, not with predictions ("retrodictions") made after a structure is known.

Russell and Barton (1993) describe one way to determine the proper target score for perfect prediction. These authors determine "expected prediction accuracy" for a given family of proteins based on a new variable which they call **conservation** (*C*) and the length class (based on number of residues) of the sequence (<=50, 51-100, 101-150 and >150), where *C* is the percentage of alignment positions sharing seven or more property states (hydrophobicity, aliphatic, etc.) as defined by Zvelebil *et al.* (1987), across all aligned sequences. Their lower limit for expected Q_3 accuracy is as low as 70%, an accuracy range achieved by several recent predictions.

One candidate for Jenny and Benner's "score for structural elements" has been developed by Rost *et al.* (1994), since "the ultimate goal is reliable prediction of tertiary . . . structure, not 100% single residue accuracy for secondary structure." Approaching the same problem as Russell and Barton (1993) in a different way, Rost *et al.* compared secondary structures of proteins with the same tertiary fold. They found an average three-state single residue accuracy of 88.4%, with a standard deviation of 9%. Accuracy for dissimilar sequences is about 35%. They propose a score, **segment overlap (*Sov*)**, which is a measure of similarity of predicted and actual segments (elements) of secondary structure:

$$\mathrm{Sov} = \frac{\sum_{,} \left(\frac{\mathrm{minov}(s_1;s_2) + \delta}{\mathrm{maxov}(s_1;s_2)} \right) * \mathrm{len}(s_1)}{N} \quad (6)$$

where *N* is the total number of residues in the protein; the numerator is summed over all segments of secondary structure; subscripts 1 and 2 are the two sequences of secondary structures being compared (1 is usually observed and 2, predicted); s_1 and s_2 are two segments, one from each sequence, that have in common at least one residue position in the same secondary structure; min*ov* is the actual overlap between the two segments; max*ov* is the total extent of either sequence, and *len*(s_1) is the length of the observed segment. That is, for a helical prediction, min*ov* is the number of residues for which both segments have an H (helical prediction) in common; max*ov* is the number of residues for which either of the two has an H. δ is an integer variable chosen to be smaller than min*ov* and smaller than one-half the length of *s1*; δ = 1, 2, or 3 for short, intermediate, and long segments. The ratio of min*ov*/max*ov* is constrained to a maximum value of 1.0.

Sov is defined such that two sequences with identical segments will have a *Sov* of 100%; in practice it can be as high as 90% for homologous sequences and is usually higher than single residue accuracy. *Sov* weighs more heavily those aspects of secondary structure that are more important in tertiary structure and Rost *et al.* suggest that both it and single residue accuracy be reported as joint measures of prediction accuracy.

Whatever the measure of accuracy used, it might be expected that when the same prediction method is assessed using the same evaluation method by different workers, the same accuracy would result. However, ". . . it can be stated unequivocally that the original claims of accuracy in the predictability of the various methods of the secondary structure of proteins have not been found to be maintained in the laboratories of others." (Fasman, 1989b).

All the above accuracy discussions assume a predictive method is being used as its developers intended. While the Chou-Fasman secondary structure prediction method can be carried out manually, it and most other methods are usually implemented in computer programs. Unfortunately, when these implementations have been checked in the case of the GOR method, a high percentage of commonly-used commercial and non-commercial algorithms are invalid (Ellis & Milius, 1994). Testing implementations of any method is strongly recommended.

[TOC] [Biophysics Textbook Home Page]

Lynda Ellis, September 29, 1998

General Empiric Methods

Empiric methods for protein structure prediction are based on experimental data. This can be statistical information on the sequences themselves, or include data on, for example, hydrophobicity of individual amino acid residues, but, since the goal is structure prediction from primary structure alone, no information based on measurements made on the protein itself. These methods are primarily statistical in nature, but also include neural networks. Most are used to predict secondary structure. Examples of tertiary structure prediction are discussed separately.

Statistical

The accumulation of known protein structures enables researchers to collect statistical information on the probabilities of various amino acids being in certain structural states within a protein. One can use these statistical probabilities to develop empirical rules for secondary structure predictions. The quality of a particular method depends both on the size and the quality of the database from which statistical information is obtained, and the way the statistical probabilities are used to develop the rules.

Over the years, several empirical statistical methods for secondary protein structure predictions have been developed. Some were designed to predict general secondary structures in all conformation states (helix, strand, turn, or random coil). Others focus only on predictions of selected states. Those techniques discussed here are based either on databases of all globular proteins, or are limited to no less than a protein structural class. Empiric statistical tecniques more appropriate to single protein families are described under Homology Modeling.

Among the most widely used methods for predicting general secondary structures are those of Chou-Fasman (1974a, 1974b) and Garnier *et al*., 1978; Garnier and Robson (1989). In the **Chou - Fasman (CF) method**, the conformational parameter of the amino acid *i* for the conformation state *X* (*X*= helix, strand, turn, and coil) from a database of a total number of *N* amino acids is defined as

$$P_{i,x} = (n_{i,x} / n_i) / (n_x / N) \quad (7)$$

where n_i,X is the number of observed amino acid *i* in the conformation state *X*, n_i is the total number of amino acid *i*, and n_X is the total number of amino acids in conformation state *X*. The amino acids are placed in order by the value of the conformational parameters for a conformation state and placed into various classes (helix former, helix breaker; strand former, strand breaker; etc.). In order to predict the nucleation, propagation and termination of helices and beta strands, as well as the presence of turns, a set of heuristic rules are included as follows:

For α helix:

(i) Helix nucleation occurs when four helix formers are found out of six residues in the sequence;

(ii) The helix continues in both directions until four helical breakers are encountered;

(iii) There are special rules for proline;

(iv) The segment is considered a helix when the average probability for helix P(α) is greater than 1.03 and P(α) > P(β), where P(β) is the average probability for β strand.

For β strand:

(i) A β strand is formed if 3 beta formers are found out of 5 residues;

(ii) The strand continues in both directions until 4 beta breakers are encountered;

(iii) The segment is a strand if P(β) > 1.05 and P(β) > P(α).

While the simplicity of the CF method makes it possible to do predictions by hand, it is usually computerized; for example, Ralph *et al.* (1987), and Prevelige & Fasman (1989). The CF algorithm not only can locate secondary structural regions but also will detect regions with the potential for conformational changes. A wide range of prediction accuracies (58% - 86%) have been reported. In general, prediction accuracy is better for a protein with a single type of secondary structure (all helix or all strand) than for a mixed type protein. Helices were found to be better predicted than strands and turns.

The **GOR** method uses information theory to predict secondary structures (Garnier *et al*., 1978). For the conformational state S_j of the *j*th residue in a sequence, the general form of information is defined by

$$I(S_j, R_1, R_2, \ldots, R_{last}). \quad (8)$$

Equation. 8 contains the information from the first to the last residue on the *j*th residue. If I>S_j has *n* possible states, then there are *n* values of information associated with each state and the highest value defines the predicted conformational state. Although Equation 8 includes every residues in the sequence, it was found that the effect on the *j*th residue is dominated by the information of residues up to eight residues distant. Therefore, Equation 8 can be approximated by

$$I(S_j, R_1, R_2, \ldots, R_{last}) \cong \sum_{m=-8}^{m=+8} I(S_j, R_{j+m}) \quad (9)$$

where $I(S_j,R_{j+m})$ represents the conformational information that the $(j+m)$th residue carries about the *j*th residue.

Four conformational states were defined in the first version of GOR method (GOR I): α-helix, β-sheet, reverse-turn, and coil. Parameters $I(S_j,R_{j+m})$ were obtained from the directional information plots of 25 proteins of known structures. For optimizing the accuracy of predictions, two more adjustable parameters, decision constant and run constant, were also introduced for each conformational state. An overall accuracy of about 60% residues correctly predicted was achieved. The apparent upper limit of accuracy in predicting secondary structures was attributed to the tertiary interactions between residues far apart in the sequence. This reasoning led to the recommendation that homologous proteins should be included whenever they are available.

Gibrat *et al.* (1987) updated GOR methodology to include a new data table, but limited to three, rather than four, secondary structure conformations, based on the directional information values from 75 proteins. Garnier and Robson (1989) expanded the Gibrat *et al.* (1987) tables to the same four state model used in GOR I. This 1989 update is called GOR II. The information-theory equations and algorithms are the same in GOR I and GOR II; they differ only in their data tables.

The combination of the single-residue information and the pair information about *j*th residue by a residue of type R_{j+m} at position $j+m$ given that a type R_j is at position j gives

$$I(S_j, R_1, R_2, \ldots, R_{n+1}) = \sum_{m=-8}^{m=+8} I(S_j, R_{j+m} \mid R_j). \quad (10)$$

Equation 10 is referred to as the GOR III equation. Theoretically, GOR III should be more accurate than GOR I or GOR II because more detailed information is used. However a comparison of the accuracies of all three GOR methods (Garnier and Robson, 1989) showed that little if any improvement could be achieved by using GOR III method for four-state predictions. In the remainder of this chapter, GOR is used to refer to the GOR II method, unless otherwise specified.

Zhang *et al.* (1992) developed yet another statistical approach, similar to GOR in that it is a pairwise methodology, but includes information on all C_n^2 possible pairs in a window of size *n*, not (*n*-1) pairs as does GOR. This method is not evaluated alone, but is one of three "experts" that form a hybrid system. The other two parts are a memory- (case-) based reasoning system and a neural network. The entire hybrid system is discussed under Neural Networks.

Other techniques or modifications of the previous techniques are useful for predicting specific structures. The helical wheel method of Schiffer and Edmunson (1967) uses hydrophobicity of residues to detect the presence of amphipathic α-helix structure. Presumed helical residues are arranged in a circle or wheel, each residue located 100 degrees from its predecessor (the angular distance separating adjacent side chains in an α-helix). Wheels of amphipathic helices would show a region with a preponderance of hydrophobes. Cornette *et al.*, (1987) expanded on this to use hydrophobicity values. Thirty-eight published hydrophobicity scales were compared for their ability to detect the characteristic period of α-helices and an optimum scale was developed using a new eigenvector method. Both discrete Fourier transform and least-squares power spectrum methods were used to find the dominant frequency of helical wheel intervals. The latter method was found to be more reliable.

Garret *et al.* (1991) modified GOR parameters to predict substates of β-residues. β-residues can be divided into two substates, internal and external, based on their distinct hydrogen bonding patterns. Internal residues are shared by two β-ladders (β-strands) while external residues belong to a maximum of one β-ladder. Two sets of GOR prediction parameters for both external and internal β-substates were developed. The overall quality of predictions is not significantly improved by the new parameters. However, the distinction between these two substates of β-residues may provide limited tertiary structural information.

The empiric, statistical methods mentioned so far in this section have made predictions based on statistics from all available (or general representative) globular proteins. As mentioned earlier, predictions based on a single protein family use different techniques and are discussed later, under Homology Modeling. There are two types of intermediate methods which will be discussed here. The first type includes predictions based on a protein's structural class. Structural-class-specific variants of the Chou-Fasman prediction tables with improved accuracy have been reported (Chou, 1989).

The technique of pattern matching is generally a "homology", rather than an "empiric statistical" method. Most examples of this are discussed under Homology Modeling, but the results for two which predict general structures in proteins of a single structural class are discussed here. Cohen and coworkers (1986, 1991) developed a hierarchical pattern search algorithm for locating turns in proteins of three structural classes (all-α, α/β, and all-β). The hierarchical order is defined by the classification of four individual patterns based on the following physical principles: local maxima in hydrophilicity; secondary structure identification and avoidance; regions containing proline; and weakly hydrophobic segments distant from well-defined turns. A 95% accuracy was achieved on a test set of proteins of known structure.

The same hierarchical pattern search approach was subsequently expanded and improved (Presnell *et al.*, 1992) to predict α-helices by dividing regular secondary structures into components and by the inclusion of a new procedure for the analysis of metapatterns. Regular helical patterns are divided into three components: the amino terminus, the core, and the carboxy terminus. The algorithm achieves a high recognition score for helix core, 95% with 10% overprediction, but appears less reliable for predicting N- and C- terminal caps, 50% with 25% overprediction. Overall, an accuracy of 71% was reported on 20 all-α protein sequences, compared to 65% using the CF algorithm, 71% from the GOR algorithm, and 78% using the neural network of Kneller *et al.* (1990). This last system, and its combination with the turn-prediction pattern-matching approach, is discussed under Neural Networks.

The second type of intermediate method determines the most similar, though not homologous, protein structures to use for secondary structure prediction, and is called the case-based method.

Leng *et al.* (1994) has developed a "case-based" secondary structure prediction method which automatically finds 55 of the most similar proteins from a structure database. "Similarity" is based on amino acid composition or sequence, but is not as close a relationship as is used in homology techniques. Then each sequence with known structure is decomposed into segments of 22 residues in length, and the segments are compared to corresponding segments in the unknown sequence, correspondance measured by a different similarity score. "Each segment of a reference protein will assign its structure to a segment of the unknown with a weight equal to the product of its similarity value and the similarity weight of the protein it comes from . . . . For each amino acid in the unknown, weights are accumulated for three classes of structures, α-helix, β-strand, and coil." The intermediate prediction for each amino acid is the class with the highest weight. A final step uses rules to fill in gaps in helices and strands and to change isolated predictions of helix or strand to coil. With a 22-residue window, this method looks at a longer segment than most. When the window was shortened, predictive accuracy decreased. Predictive accuracy of this method is in the 70% range, which puts it in the high end for secondary structure prediction in non-homologous proteins.

Neural Networks

A computer-based neural network is a computer program that can, wth training, detect correlations in data and learn to recognize patterns. In principle, neural networks can detect second- or higher-order correlations in data; therefore, they can be more powerful than methods based on standard first-order statistical treatments. In this section, we shall briefly describe how a neural network works and how one can be used to predict protein structure.

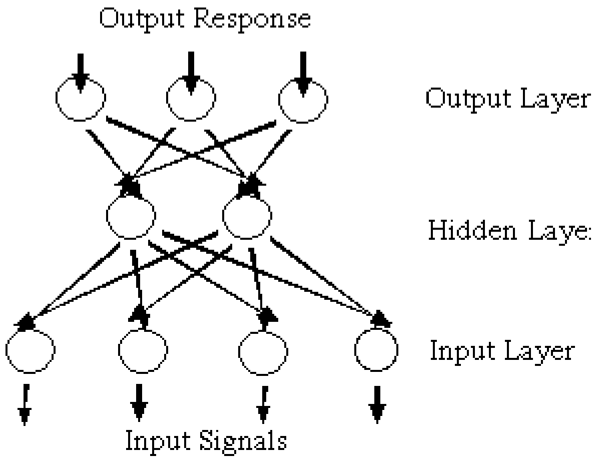

Figure 1. Typical neural network architecture.

A typical neural network contains an input layer of units which receives input signals, an output layer of units which outputs structure predictions, and zero, one, or more hidden layers in between the input and output layers. A network without hidden layer is called a **perceptron**, and can only detect first-order correlations between input signals and output responses. Networks with hidden layers can extract higher-order features. The units of the network are connected to one another with a real number as a weight. These weights determine the behavior of the neural network. A neural network is trained to recognize patterns by changing the weights between various units. When using a neural network to predict protein structure based on primary sequence, several decisions have to be made: (i) the method of encoding a sequence; (ii) the number of input units; (iii) the number of hidden layers; and (iv) the choice of learning algorithm (Hirst and Sternberg, 1992). The typical learning procedure used in protein secondary structure predictions is the **back-propagation of errors algorithm**. It maps input to output by adjusting the connecting weights to minimize the difference between the computed and the desired output unit values.

An early work using a neural network for predicting the secondary structure of globular proteins (Qian and Sejnowski, 1988) used a segment of protein structure of fixed size as input. The length of the segment is called the input window size. A sequence of 21 binary numbers are used to encode each amino acid residue, 20 for identifying the amino acid (19 zeros and a single 1) and 1 for regions between proteins. For an input window of *N* amino acids, the input layer of the neural network needs at least *N* group of units with 21 units in each group. The output layer of the neural network consists of 3 units corresponding to 3 types of secondary structures: α-helix, β-sheet, and coil. The network maps the input sequence of amino acids and outputs the state of the middle residue in the sequence. The input window moves through the protein, while the network predicts the secondary structure of one amino acid at a time.

A database containing 106 proteins was used in the study. A subset of these proteins was used for training the network and the remaining proteins were used for the testing. Since the results were highly sensitive to homologies between proteins in the training and testing set, care was taken in the selection of proteins so that no homologies were present in the training set.

Predictions using various input window sizes show a maximum success rate at around 13. Reduced performance at smaller window sizes probably results from lack of information on residues outside of the window. The performance deterioration associated with larger window sizes was attributed to the interference from irrelevant input groups (residues).

Networks with various numbers of hidden units (0 - 60 in one hidden layer) were trained and tested in the study. The results show that the testing success rate was almost independent of the number of hidden units, suggesting that only first order features were present in both the training and the testing protein sets. However, the learning rates of the training set became slower as the number of hidden units decreased.

Improved performance on predicting α-helix and coil regions (but not β-sheet regions) was achieved when cascaded networks were used. The input to the second network was sequences of outputs from the first network, 13 groups of units with 3 units per group. Each group contains the secondary structure assignment derived from the first network. The improvement of performance was attributed to the fact that the second network joins short fragments of secondary structure and eliminates isolated assignments from the predictions of the first network.

The prediction accuracy on the testing set was found to increase with the increasing size of the training set as a result of better generalization. However, the success rate approaches to a limiting value beyond certain size of training set. This implies that adding more protein structures to the training set will not result in significant improvement for non-homologous proteins. A number of modifications to the basic network did not lead to performance improvements. These included modifying the input by incorporating biophysical properties (such as charge, size, hydrophobicities, etc.) of the amino acids. More detailed secondary structure classification (three types of helices, two types of β–structures, two types of turns, and coil) also failed to increase the accuracy of predictions. Finally, various modifications to the network architecture were carried out with little or no performance improvement (Qian and Sejnowski, 1988).

Qian and Sejnowski (1988)compared their prediction results to those from early versions of statistical methods including those of Chou-Fasman, GOR and Lim using a non-homologous testing set of proteins. The Q_3 value of the neural network was found to be about 10% better than those from the three statistical methods. The Mathews correlation coefficients of the neural network are also better (>30%) than those of the statistical methods. However, it should be pointed out that improvements on the statistical methods have been made since 1988. The current level of performance of these methods are comparable to that of Qian and Sejnowski's neural network.

Neural network methods have also improved and a more recent comparison between them and statistical methods for structure prediction is of interest. The task chosen was identifying ATP-binding motifs (Hirst and Sternberg, 1991). The feed-forward neural network had two layers, the input window size was 17 residues with 20 units for each residue, and the output layer was a single unit to predict whether the sequence would bind ATP or not. The statistical method used in the comparison was a motif-searching program which measures the similarity between two sequences by defining a pattern using a set of aligned sequences and computing the score between the test sequence and each of the pattern-defining sequences. The average score from the summation over all the sequenes reflects the degree of homology between the test sequence and the set of pattern-defining sequences.

For the comparison, 193 ATP-binding proteins and 156 ATP-non-binding proteins were chosen from the SWISSPROT database. For testing the neural network, one sequence was removed out of the 349 sequences, and the network was trained on the remaining 348 sequences. Prediction was performed on the removed sequence. The procedure was repeated for each sequence so that the network was tested on all sequences. For the statistical method, one of the 193 ATP-binding proteins was removed for testing, and a binding pattern was defined using the remaining 192 sequences. Prediction was then carried out on the removed protein, and the procedure was repeated for each of the ATP-binding proteins. In the case of ATP-non-binding proteins, an ATP-binding pattern was defined on all 193 ATP-binding proteins and then a score was calculated for each of the 156 ATP-non-binding sequences. The methods were essentially identically accurate; the neural network correctly predicted 78% of the 349 sequences, whereas 80% of these sequences were correctly classified by the statistical method (Hirst and Sternberg, 1991).

Returning to predicting general secondary structure, Rost and Sanders (1993a) developed a three-level percepton (neural network with no hidden layer) which uses multiple sequence alignments as input instead of single sequences with a window size of 13 consecutive residues. This inclusion of protein family information increases the prediction accuracy by 6-8 percentage points. A combination of three levels of networks results in a three-state accuracy of 70.8% for globular proteins. Another methodology, inductive logic-based machine learning, differs from either statistical or neural network methods, yet is empirically based. Muggleton *et al.* (1992) used such a system trained on 12 all-α proteins to predict residues in helical structures for four all-α proteins not in the training set. The predictions had an accuracy of 81%, compared to 72% using GOR. One of the 12 sequences was mistakenly included though it had 44% sequence identity to another. However removal of one of these two homologous proteins did not reduce the accuracy of the predictions for all the four all-α proteins (Muggleton *et al.*, 1993).

Rost and Sander (1993b) studied the same four sequences using two neural networks. One network was trained on two states (helix, non-helix) using Muggleton and co-worker's training set; the other one, a general secondary structure prediction network (Rost and Sander, 1993a), was trained on three states (helix, strand, loop), using 130 proteins of all structural classes, and then evaluated on two states. An overall two-state single residue accuracy of over 80% was obtained by either network, and the authors concluded "there is no practical advantage in training on two states, especially given the added margin of error in identifying the structural class of a protein."

Holley and Karplus (1989) also explored the application of neural network for protein secondary structure prediction. Their network consists of an input layer of 17 units (residues), compared to a 13 residue window used by Qian & Sejnowski (1988), a hidden layer of 2 units, and an output layer of 2 units - one for helix and the other for sheet. A back-propagation learning algorithm was used in training of the network. 62 proteins were used in the study with the first 48 proteins for training and the remaining 14 for testing. An overall predictive accuracy of 63% for 3 states (helix, sheet, and coil) was achieved. This accuracy is about 10% higher than those of the statistical methods of the time. Similar to the findings of Qian & Sejnowski (1988), no accuracy loss was noted when the hidden layer was removed, implying that only first order correlations were extracted.

Attempts were made to improve the results by adding periodic sequence information to the neural network and by subdividing proteins into all-α, all-β, α/β and "other" classes (Kneller *et al.*,1990). Based on the fact that α-helices often have a hydrophobic and a hydrophilic side, additional inputs representing the hydrophobic moments defined by Eisenberg *et al.* (1982) were introduced to the network, resulting in limited prediction accuracy improvement. On the other hand, the inclusion of a special unit that looks for complementary charge pairs between residues *i* and *i+4* (a common feature of helices) failed to show any improvement. The authors also subdivided their database of 105 proteins into four **structural classes**: all-α (22), all-β (24), α/β (20), and "other" (39). (Methods to predict structural class are discussed later.) Training and testing were performed on each individual class of proteins. Compared to the results of Qian and Sejnowski, enhanced accuracies were obtained for all-α and all-β classes, while no improvement was noted for the α/β class. The all-α protein class also showed a strong correlation between prediction accuracy and sequence identity. No such correlation was found for the all-β class.

This work continued, and was combined with the pattern matching approach of Cohen *et al.* (1986, 1991) described earlier, in the development of MacMatch (Presnell *et al.*, 1993), a package for secondary structure prediction. Turns are predicted by pattern matching, with three different patterns for all-α, all-β and α/β structural classes. If structural class is unknown, the α/β class may be used as default, or it can be predicted from amino acid content as discussed later in this section. Tested and trained neural networks, which use either general or structural-class-specific weights, predict helix and strand. The authors report neural network predictive accuracies of 79, 71, and 64% for all-α, all-β and α/β proteins of known structure.

As mentioned earlier, a second combined predictive methodology was developed by Zhang *et al.* (1992). Instead of usig discrete, separate methods for different aspects of structure, this is a hybrid system, with three structures prediction "experts," statistical, memory-based, and neural network. The results of each expert are examined by a combiner to make the final prediction. All three methods used a window size of 13 residues to predict the structure of the center residue. The statistical method has been described earlier. The memory-based method is similar to the case-based method of Leng *et al.* (1994) described above, except that cases are segments only, 22 most similar sequences were used instead of 55, and a different similarity matrix is used.

The neural network is a back-propagation network with one hidden layer with only two units. The combiner is also a one-hidden-layer neural network, with 30 units in the hidden layer, and is trained; that is, it learns the best way to combine results. The three-state accuracy of the hybrid system is 66.4%, better than the C-F, GORIII, Qian and Sejnowski (1988), and Holly and Karplus (1989) methods described earlier. The fact that 20% of the time all three systems produced the same, wrong, prediction, suggests an upper bound on the accuracy of empirical methods at the present time (Zhang *et al.*, 1992).

The inclusion of tertiary interaction information to the input of the neural network could, in principle, increase the secondary structure prediction accuracy. Vieth and Kolinski (1991) added the information of contacts between the central residue of the 13-residue input window and other residues in the sequence to the input layer of a network similar to that of Qian and Sejnowski (1988). In addition, three filtration procedures were developed to deal with the problem that the network result for the entire protein often contains unphysical sequences of structure assignments: (i) terminal residues (C- and N-terminus) are given coil assignments; (ii) structures of the type HXH or HHXHH are replaced by HHH or HHHHH, where H represents a helical assignment and X is for coil or β-strand; and (iii) separated α or β assignments of one or two residues in length are replaced by coil ones. A set of 39 proteins from Brookhaven Protein Data Bank were used: 31 of them for training the neural nets and 8 for testing. The combination of the inclusion of tertiary interactions and the filtration rules was found to increase the prediction accuracy by 3-5% compared to a network without these treatments; however, adding tertiary interactions means this is not prediction based on primary structure alone.

Vieth and co-workers (1992) continued by explorig the advantages of cascade network and filtration rules. They developed a complex network containing four cascaded networks with two simple networks per cascaded one and three sets of filtration rules. The first cascaded network was used for assigning the structural class of the protein (α, β, or α/β). The other three cascaded networks were trained individually for three specific classes. Once the structure class of the protein was determined, the sequence was sent to one of the specific cascaded networks, followed by a filtration for final structure assignment. When only two hidden units were used in each simple network, it was found that the complex network increased the prediction accuracy by 2-4% compared to a simple network. However, little improvement was obtained when the number of hidden units increased to 40.

Tertiary Structure

The previous empirical statistical and neural network methods predict secondary structure. A more ambitious goal is to use similar techniques to predict tertiary structure. A step toward complete tertiary structure prediction is the prediction of a protein's structural class. Structural class is usually defined as one of four different groups of protein folds, based on the predominant secondary structure: all-α, all-β, α/β (α alternating with β), and α + β (one or more all-α and all-β domains or regions). For completness, a fifth small/irregular class is sometimes included in the classification, though not included in predictive methods.

While circular dichroism and other techniques can predict structural content and class when measurable quantities of protein exist, it is more difficult to do so when only a protein sequence is available. Cohen *et al.* (1993) suggest the use of amino acid composition to predict structural class in those cases. Chou (1989) developed an algorithm that can assign the correct structural class to a protein based on its amino acid composition with 80% accuracy. In this method, the average mole percent amino acid composition of representative proteins is four structural classes (all-α, all-β, α/β, and α + β) is calculated as shown in Table I. The amino acid composition of the unknown protein sequence is compared to each column of the table, the absolute values of the differences found for each amino acid type are summed, and the protein is assigned to the structural class to which it has the least total compositional difference.
[Table I]

Zhang & Chou (1992) expanded on Chou's method to achieve 83% accuracy. Metfessel *et al.* (1993) compared two different neural networks and a statistical approach for predicting structural class based on amino acid composition and hydrophobic residue pattern frequency. They similarly obtained a predictive accuracy of about 80%, and the differences between the three different methods were not statistically significant. As mentioned above, Vieth *et al.* (1992) used a cascaded neural network to predict structural class (α, β, or α/β), with 79% accuracy, as part of their secondary structure prediction method. Muskal & Kim (1992) have developed a neural network which predicts secondary structure content (and thus can partition proteins into all-α, all-β, and mixed α+β and α/β classes), given a protein's amino acid composition, molecular weight and the presence or absence of heme, with greater than 90% accuracy. However, heme content is not known from primary sequence.

Wilcox *et al.* (1990) and Xin *et al.* (1993) predict complete tertiary structure using a back-propagation neural network on a supercomputer. They began by using a small training set of 15 proteins, which was later extended to 20 proteis, each with less than 133 residues. Their network had one hidden layer. About twice as many hidden units as training items are required for optimum convergence. They use the entire hydrophobicity-coded sequence as input and produce a distance matrix as output. The trained network can predict reasonable structures (average RMS residual error ca. 0.05) for novel proteins as long as there are at least two respresentatives of that protein's family in the training set.

[TOC] [Biophysics Textbook Home Page]

January 4, 1999

C. Homology Modeling

Definitions

The terms "homology" and "homology modeling" can have varying definitions. Some would restrict the phrase "homologous proteins" to only those sequences that share a common ancestor (Doolittle, 1987). Since it is difficult to prove or disprove ancestry for proteins which have less than about 30% residue identity, one alternative, the one used here, more broadly defines the phrase to refer instead to structural similarity.

The phrase "homology modeling", defined most strictly, means predicting the tertiary structure of an unknown based on the known coordinates of a protein to which it is homologous (has a high degree of sequence identity/similarity). Here again we will use a broader definition to include, in addition, predictions of secondary structure carried out in a similar fashion, the development of consensus patterns, and other predictive methods developed for a single protein family.

A more important concept to define is "similarity." That is, how much similarity to a known sequence do you need, in how long a sequential run of residues, to accurately predict structure? To gain insight into the first question, Hilber *et al.* (1993) studied pairwise superpositions of a large number of known structures from different conformational and functional classes with various degrees of homology, and suggested the following relationship between sequence homology and structural differences:

(i) The size of the common core region decreases with decreasing sequence identity. Pairs with identity greater than 50% have over 90% of their residues in structurally conserved regions. If sequence identity drops below 20%, the common cores contain about 65% of the amino acids.

(ii) The overall r.m.s. difference of corresponding α-carbon atoms increases as the sequence identity decreases, ranging from 0.32Å for identity near 100% to 3.66Å for about 20% identity.

(iii) Structurally divergent regions (loops, turns) with more than 50% sequence identity have similar conformational structures. Greater structural deviations may occur for homologous loop regions with lower degrees of sequence identity.

(iv) Decreasing sequence identity correlates with increasing numbers of insertions and deletions. A maximum of 16 was observed for a case with about 20% identity. On the other hand, virtually no insertions are needed when the identity is beyond 60%.

Although there are exception to these general relations between sequence homology and structural differences, the above observations provide some tools for the assessment of accuracy at a given level of sequence homology.

How large a run of similar or identical residues can be used to imply structural similarity? Kabsch & Sander (1984) demonstrated that even exact sequence identity, in small enough segments, gives no indication of structure, by providing examples of sequentially identical pentapeptides that adopted different structures in different proteins. Wilson *et al.* (1985) extended this to hexapeptides. However Cohen *et al.* (1993) reexamined hexapeptides and found that, within a protein structural class (all α, all β, α/β, α+β), the structural similarity of sequentially identical hexapeptides usually is preserved.

This last study encourages the development of structural-class-specific secondary and tertiary structure prediction algorithms. Chou (1989) describes such structural-class variants of the Chou-Fasman secondary structure prediction tables and reports improvements in predictive accuracy if these are used where structural class is known, instead of the standard C-F tables based on all proteins.

Consensus Patterns

Closely related to predicting protein structure through runs of identical sequence is the use of patterns of consensus sequences. These are patterns of amino acid residues, optionally including alternative residues at selected positions, variable gaps in sequence, and predicted secondary structure.

As an example, Boldt and co-workers (1995) recently reported a revised consensus pattern for extradiol dioxygenases:

```
(G,T,R,N)X(H,A)X{7}(L,I,V,M,F)YXX(D,T,E,A,N)PX(G,P)X{2,3}E
```

The first 11 residue positions of this 21- or 22-residue-long pattern can be read as: A glycine, threonine, arginine or asparagine residue followed by any residue followed by a histidine or alanine residue followed by any seven residues followed by a leucine, isoleucine, valine, methionine or phenylalanine residue. The "`X{2,3}E`" at the end of the pattern specifies that the terminal glutamate residue is preceded by two or three residues of any type.

As mentioned in the Introduction, patterns of this type are evaluated by their sensitivity (measure of how frequently they occur in the proteins which are known to share the given structure or functionality), and specificity (measure of how frequently they do not occur in proteins which are known not to have the structure or functionality). The extradiol dioxygenase pattern has 100% sensitivity and specificity, occuring in all known members of a certain class of extradiol dioxygenase sequences and in no other sequences in the 67,423-member PIR40, or the 36,000-member Swiss-Prot 28, protein sequence databases. This pattern could be used to predict extradiol dioxygenase structure and activity if it were found in a protein for which only sequence was known. No tertiary structure is known at present for any member of the family. Thus, though the pattern does include several residues thought to be involved in metal binding, it is not certain what structural feature(s) it specifies (Boldt *et al*., 1995).

While such consensus patterns can be developed from **multisequence alignments**, most are based on a structurally and/or functionally important part of a known tertiary structure, and are used to predict this structural element. The PROSITE database (Bairoch, 1993) is a compendium of consensus sequence patterns. Various search engines can accept a sequence as input and search PROSITE to discover if the sequence contains any of PROSITE's patterns.

Most commercial computer-based protein sequence analysis packages include modules to search protein databases for primary sequence patterns as complex as the one for extradiol dioxygenases. Cohen and coworkers (1991) have developed a more elaborate pattern specification language, PLANS, which, as discussed under Empirical Statistical Methods, they have used to predict turns and alpha helices. PLANS can be used for the more specific patterns discussed here; Cohen and coworkers use of it is more general since the structures they predict are found in a structural class, not a specific protein family.

As mentioned earlier, consensus patterns can optionally include predicted secondary structure. The Ariadne pattern specification language (Lathrop *et al*., 1987, 1993) is one that facilitates this. As an example, part of an Ariadne pattern for the thioredoxin redox motif is:

```
(b-strand
(* :gap-min -9 :gap-max -1 :gap-max-overrun 0)
aliphatic
```

This can be read as: "Search for residues that are predicted to be in β-strand secondary structure. When these are found, go to the end of the predicted strand and search for the second pattern element (an aliphatic residue). The gap required between the first and second element is -9 to -1. A gap size of zero would have the second element directly follow the first. A gap range of -9 to -1 states that the second element is found between the 9th to the first residue from the C-terminal end of the first element." The complete pattern was found in the ORF3 protein in the *mer* operon of *Staph. aureus* and this hypothetical protein was predicted to be structurally similar to thioredoxin (Ellis *et al.*, 1992).

Other Methods

Later in this section we will see how multiple sequence alignment can improve secondary structure prediction, but the reverse is also true: secondary structure prediction can improve alignments. For example, gaps and insertions may be placed in regions that do not have a high probability of secondary structure. An extension to the Pattern-Induced Multiple Alignment (PIMA) program (Smith & Smith, 1992) can employ secondary-structure-dependent gap penalties given the tertiary structure of one or more members of the same family. The use of secondary structure information can significantly improve the accuracy of aligning structure boundaries.

If we do not use knowledge of sequence similarity, it is a truism that secondary structure prediction is no better for homologous than for inhomologous sequences. For example, if the Chou-Fasman and GOR prediction methods are used on four highly homologous ribonuclease structures, only three of the nine helices and strands are correctly predicted by both methods and in all sequences (Cohen and Cohen, 1994). However**,** secondary and tertiary structure are better conserved than amino acid sequences in homologous proteins. One method to use known structure in structure prediction is to align homologous sequences, optionally make three-dimensional models, and predict the structure of the unknown sequences based on the known structures. The serial publication Protein Profiles, described in more detail under "Resources", includes in each issue multisequence alignments of a given protein family.

In a good example of homology modeling in the strict sense, 17 thioredoxins and seven thioredoxin-like domains from protein disulfide isomerase and its homologs were aligned. Models of tertiary structure were constructed and secondary and tertiary structure was predicted based on one known tertiary structure (Eklund *et al.*, 1991). With such close homologs, additional tertiary structures may not be necessary. The structure predictions made in this manner may be off by one or two residues at the start or end of an element of secondary structure, but the overall **fold**, or backbone tertiary structure, of each member of the family is expected to be very similar. However this method works less well as sequence similarity decreases. For example, Ellis *et al.* (1992) made secondary and tertiary structure predictions for several sequences, including the DsbA protein, based on sequence alignments and pattern matches with thioredoxin. Later, an x-ray crystal structure of DsbA indeed showed it to have structural similarity to thioredoxin. However, the prediction was incorrect in the C-terminal half due to a 80-residue insertion in the DsbA sequence not found in the alignment (Martin *et al.*, 1993).

In a second example, Boniface and Reichert (1990) predicted thioredoxin structure and functionality in follitropin and lutropin, though those hormones had little sequence identity to the thioredoxin family. A second prediction of Ellis *et al.* (1992) was that these two protines and other members of the glycoprotein hormone family do not contain structural analogs of the thioredoxin redox motif. This last prediction has recently been confirmed (Lapthorn *et al.*, 1994; Wu *et al.*, 1994).

Local homologies have been used in more elaborate ways to improve secondary structure predictions. The basic assumption once more is that similar peptide sequences have similar secondary structure tendencies. Zvelebil *et al.* (1987) used multisequence alignment of homologous proteins to predict secondary structures based on the GOR I method (Garnier *et al.*, 1978). Using the observation that insertions and sequence variations tend to occur in loop regions, the algorithm first aligns a family of sequences and obtains a value of the extent of sequence conservation at each position. Then this value is used to modify the GOR prediction. Predictions performed on 11 proteins with a known secondary structure and more than four homologous sequences show an average improvement of 9% in accuracy over the GOR I method. Up to 4% improvement can be found by simply averaging the predictions at aligned positions (Zvelebil *et al.*, 1987).

A more detailed study using seven protein families and several multiple alignment programs showed that a mean increase of 6.8% in accuracy could be achieved when the minimum sequence identity between all the members in a group of homologous proteins is greater than 25% (Levin *et al.*, 1993). The increase in prediction accuracy was attributed to the extended information provided by very distantly related sequences.

As described under Neural Networks, Rost and Sander (1993a) also predict secondary structure using information from multisequence alignments rather than individual sequences. This was used as input to a neural network, and three-state accuracies of greater than 70% were achieved. Such accuracy is comparable to that of physical measurements such as circular dichroism specroscopy.

Boscott *et al.* (1993) improved the secondary structure prediction step in homology modeling in a different manner. Each of four algorithms is used to predict the structure of a protein with known structure homologous to the unknown protein. Then the customized weighted average structure prediction (WASP) algorithm which best predicts the structure of the known homolog is used to predict the secondary structure of the unknown protein. They reported an improvement ranging between 3% to 7% over the GOR method.

Donnelly *et al.* (1994) looked specifically at helix prediction in homologous protein families using multisequence alignments, environment-dependent (but family-independent) substitution tables, Fourier transform methods, and helix capping rules. They tested the method on four protein families: homeodomain, indoleglycerol phosphate synthase, insulin, and cytochrome c. Averaging the results over all four families, they can correctly predict 79% of the residues in helices, compared to 69% using GOR, and only overpredict 12% of non-helical residues as helical, compared to 35% using GOR. The method reliabily predicts the correct number and approximate position of the helices. It also reliably predicts the internal face of each helix, thus can be used for predicting their tertiary arrangement.

Applications

With the wide range of techniques available to homology modelers, it is interesting to examine representative predictions, the methods used by each, and how successful were their predictions. Steven A. Benner and coworkers are among the most prolific of recent homology modelers. Protein families or domains they have studied include protein kinases (Benner & Gerloff, 1991), Src homology domain 3 (Benner *et al.*, 1993), nitrogenase MoFe proteins (Gerloff *et al.*, 1993a), hemorrhagic metalloprotease family (Gerloff *et al.*, 1993b), and the pleckstrin homology domain (Jenny & Benner, 1994b). They have also developed methods to predict interior and surface residues (Benner *et al.*, 1994), which in turn have been used to predict interior and surface residues which in turn can be used for secondary structure prediction. For example, 3.6 residue periodicity in surface and interior assignments can predict a surface helix, and consecutive interior assignments can predict interior β-strands. Their technique can be summarized as (Benner, 1992): get 10 to 20 homologous sequences, some pairs of which have high (70-80%), some moderate (40-50%), and some low (~30%) sequence identity. Align the sequence. Assign surface, interior, active-site and parsing (those that lie between secondary structure elements) residues. Assign secondary structure. Build models of tertiary structure.

Since Benner and co-workers emphasize predicting unknown structures, confirmation of their predictions must wait on experimental verification. This was forthcoming for the protein kinase prediction (Benner and Gerloff, 1991), and ". . . Benner and Gerloff's prediction of the core secondary structure was much better than that achieved with standard methods." (Thorton *et al.*, 1991). The Src homology domain 3 secondary structure prediction correctly predicted four of the five secondary structural elements for a "per segment" accuracy of 80% (Benner and Gerloff, 1993).

A number of other groups have made homology-based predictions. A few representatives are briefly mentioned here, including such diverse protein families as: tryptophane synthetase (Crawford *et al.*, 1987); aminoacyl-tRNA synthetase (Jentoft *et al.*, 1992); creatine kinase (Mühlebach *et al.*, 1994); matrix metalloproteinase (Hodgkin *et al.*, 1994); protein serine/threonine phosphatase (Barton *et al.*, 1994); and flavodoxin (Caldeira *et al.*, 1994). The most recent of these predictions have yet to be verified.

Crawford and co-workers (1987) used multisequence alignment, and CF, GOR, average hydropathy, and chain flexibility calculations to predict the secondary and tertiary structure of the α-subunit of tryptophan synthetase. It was predicted to have eight repreated β-loop-α-loop motifs, and an α/β barrel tertiary structure. The prediction agreed quite well with x-ray crystallography, with the sequences of all nine helices and all but β7 and β8 of the eight strands, sharing overlapping residues.

Burbaum and co-workers (1990) reviewed work in predicting the structure of a subclass of the aminoacyl-tRNA synthetase family. The members of this subclass, which synthesize bacterial Arg-, Gln-, Glu-, Ile-, Leu-, Met-, Trp-, Tyr-, and Val-tRNAs, all contain a signature consensus sequence. In the two cases where complete or partial structures are known, the residues in the signature sequence are superimposable. They thus probably form the same three-dimensional structure and have the same function in most, if not all, of the members of this subclass. While known structure can often be used to predict important residues for a homologous sequence with unknown structure, they offer an example where the significant loss of activity in a mutation at Gly94 in Ile-tRNA synthetase (with no known structure) lead to a prediction of functional importance for the homologous Ala-50 residue in Met-tRNA synthetase (with known structure).

Jentoft and co-workers (1992) tested the prediction that mammalian dihydrolipoamide dehydrogenases have structural similarity to the known structure of human glutathione reductase. Conservation of polar and non-polar groups in predicted secondary structural elements, conservation of active site residue functionality, and conservation of residues at the predicted dimeric interface all support this prediction. They next created a model tertiary structure for human dihydrolipoamide dehydrogenase based on human gluathione reductase, and used this to predict a highly polar, negatively charged active site for dihydrolipoamide dehydrogenases.

Mühlebach and co-workers (1994) have made structural predictions for creatine kinase (CK) isoenzymes. They defined a structurally important "CK framework" consensus pattern of the most conserved sequence blocks or regions, and "diagnostic blocks" which serve as signature sequences for each CK isoenzyme subfamily. Only the first block of the CK framework is missing in the invertebrate guanidino kinase sequences, leading them to speculate that this block determines the guanidino substrate specifically.

Hodgkin and co-workers (1994) made successful structure prediction of the catalytic domain of matrix metalloproteinases, using the methods of Zvelebil *et al.* (1987) for conformational propensity, combined with other methods for surface probability and residue conservation or variation. Although their prediction was for an entire protein family, when it was applied to one structure that was later determined, all strands and helices, except one, were correctly predicted, and the predicted ends of secondary structural elements were off by no more than 3 residues, compared to the crystal structure of Lovejoy *et al.* (1994).

Barton and co-workers (1994) used multisequence alignments to guide secondary structure predictions for serine/theronine-specific protein phosphatases, which lead to prediction of two domains, one with a β-sheet with flanking helices, and the other predominantly helical, and, coupled with similarity to *Escherichia coli* diadenosine tetraphosphatase, prediction of a phosphate-binding site at the N-terminus α-helix.

Caldeira and co-workers (1994) modeled the *Desulfovibrio desulfuricans* flavodoxin on the highly similar (49% identity) *D. salexigens* structure. Since these two proteins are highly similar and can be aligned with no gaps or insertions, they kept backbone coordinates the same, changed side-chains where needed and used a four-step energy minimization technique (described under Molecular Modeling) to change bond angles and lengths to minimize the energy of the molecule. The x-ray crystal structure of this molecule is currently being determined.

The examples should end on a cautionary note. Lustbader and co-workers (1993) predicted the tertiary structure of human chorionic gonadotropin using empirical and molecular modeling techniques including chemical studies on homologous members of its protein family (thus not based on primary sequence alone). These predictions proved to be incorrect; this class of glycopeptide hormones was found to have an intricate "cysteine-knot" fold (Lapthorn *et al.*, 1994; Wu *et al.*, 1994). This last example reminds us that tertiary structure prediction is still not an exact science at this time. However homology-based techniques have had greater success compared to other methods of structure prediction. This has been succinctly summarized by Rees (1990): "The answer to the question 'Structure from sequence?' is 'Not yet, unless you know what the structure looks like.'"

[TOC] [Biophysics Textbook Home Page]

November 9, 1998

Molecular Modeling

While the preceeding sections have focused on empirical protein structure prediction, using statistical or neural network or homology techniques, a more theoretical approach seeks to predict less empirically, from "first principles." The fundamental assumption is that the native structure of a protein corresponds to the conformation which has the lowest energy. Then tertiary structure of a protein can be predicted if the energy of the protein system can be calculated and the conformation associated with the lowest energy can be found.

The energy of any molecule can, in principle, be calculated using quantum mechanics. In reality, the high cost of computation prohibits the application of quantum mechanics on all but the simplest macromolecules. Instead, approximations of **potential energy functions** have been extensively used in the calculations of protein conformation energy. The approximations used often limit their accuracy. Nor is discovering the lowest energy conformation without problem. The large number of degrees of freedom in protein systems makes an exhaustive conformation search impractical. For instance, it would take approximately 108 hours of supercomputer time to simulate the folding of a protein starting from an extended polypeptide chain in solution (Karplus and Petsko, 1990).

In an attempt to partition the problem into smaller pieces, workers have examined important partial structures, and more simplified models of complete structures. Much effort has been devoted to the development of better potential energy functions for calculating the energy of the protein system, more efficient methods for searching for the lowest energy conformation, and simpler models of protein structure. In this section, we will briefly introduce the basic concepts and major techniques used for predicting tertiary protein structure based on the minimum energy assumption and survey current research in the area.

Molecular Mechanics

Molecular mechanics is a computational method designed to give accurate structures and energies of molecules. It treats molecules as collections of masses that are interacting with each other via harmonic (or more complicated) forces between bonded atoms and via van der Waals and electrostatic forces between non-bonded atoms. Mathematical functions of the atomic coordinates (called potential energy functions) are used to describe these interactions. Various parameters derived from experimental observations are included in the potential energy function, also known as the **force field**.

Although the basic ideas behind molecular mechanics can be traced back to D.H. Andrews (1930), practical procedures using these basic ideas were first implemented in the 1970's. One of the most widely cited implement was introduced by Burkert and Allinger (1982). The basic idea of molecular mechanics is that simple molecules have "natural" bond lengths and bond angles. Any structural deviation from such "ideal" molecular geometry will result in an increase in potential energy. One of the fundamental assumptions of molecular mechanics is that the total potential energy of a molecule can be divided into several parts. A typical potential energy function form widely used for proteins is (Brooks *et al.*, 1983):

$$E(R) = \frac{1}{2}\sum_{\text{bonds}} k_b (b - b_0)^2 + \frac{1}{2}\sum_{\text{angles}} k_\theta (\theta - \theta_0)^2$$
$$+ \frac{1}{2}\sum_{\text{torsional}} k_\omega [1 + \cos(n\omega - \delta)] + \sum_{\text{non-bond}} \left(\frac{A}{r^{12}} - \frac{B}{r^6} + \frac{q_1 q_2}{\varepsilon r}\right) \quad (12)$$

where *E(R)* is a function of the coordinate set, *R*, of all the atoms in the system. The first term corresponds to a Hooke's law description of bond stretching. The second term is a similar approximation to the energy of bond angle bending. The parameters k_b and k_θ are force constants that determine the flexibility of the bonds, b_0 and θ_0 are natural bond length and bond angle, while b and θ are the actual bond length and bond angle. The third term accounts for the energy associated with torsional angle rotations. The last term represents the non-bonded interactions between two atoms separated by distance r. It has three parts: the first two are the Lennard-Jones 6-12 potential which includes both short-distance repulsive and long-distance attractive interactions, and the last one corresponds to the electrostatic energy where q_1 and q_2 are the charges on atoms 1 and 2. Parameters A and B depend on the atoms involved and ε is the dielectric constant of the medium.

The preceding force field discussion is based on the assumption that a force field for macromolecular systems can be treated as a combination of force fields determined for many smaller molecular systems. Such force fields have been extensively used in the molecular modeling of macromolecutes during the last two decades. More recently, force fields contructed from known protein structures have gained increasing attention. Sippl (1993) describes the physical principles behind these so-called knowledge-based mean fields and discusses applications of these fields.

Two physical principles are used in the knowledge-based approach: (i) at equilibrium, the native state of a protein system has the global minimum free energy; (ii) the distribution of molecules among the microscopic states is governed by **Boltzman's distribution law**. Know tertiary protein structures are used to determine mean force energies of intramolecular amino acid pair interactions as a function of the distance between atoms. Protein-solvent interactions are calculated in a similar fashion. The predictive power of these knowledge-based mean force fields was tested using 157 proteins of known structure. They successfully indentified 94% (148) of the native conformations. Possible applications of the knowledge-based mean fields include the validation of experimentally determined protein structures, database searches for identifying native-like sequence structure pairs, sequence structure alignments, and conformation calculatios from amino acid sequences (Sippl, 1993). Other potential function have also been used. Examples of programs used for the modeling of biomolecular systems which incorporate various different force fields include: AMBER (Kollman, 1991), DISCOVER (see Resources), and Empirical Conformational Energy Program for Peptides (ECEPP) (Nemethy *et al*., 1992).

Once a potential function is chosen, another factor to consider in a molecular mechanics simulation is how the minimum energy conformation is determined. The landscape of a potential energy surface as a function of the coordinates of all the atoms in a system has many peaks (local maxima) and valleys (local minima). Each valley corresponds to a stable or semistable state of the system. For a protein, the structure associated with a stable state is called a conformation. Therefore, conformations can be found by locating the local minima on a potential energy surface. The computational method that starts with a set of atomic coordinates of the system and finds a nearby potential energy local minimum is call **energy minimization**. Various energy minimization methods are available. The methods using the first-derivatives of the potential energy function are usually less computationally intensive, while higher accuracy can often be achieved by using the methods involving both the first- and second-derivatives.

As mentioned before, the fundamental idea of predicting the structure of a protein using molecular modeling relies on the assumption that the conformation with the lowest potential energy is the native conformation of the protein. Therefore, the task of finding the native structure of a protein becomes the search for the global potential energy minimum. Most energy minimization methods can search only in a "downhill" direction and are unable to overcome energy barriers.

Various conformational search algorithms have been developed to sample a large area of the conformational space in order to locate the global minimum; for instance, Ferguson and Raber (1989) developed a random incremental pulse search algorithm. However, the number of minima increases dramatically as the size of the protein increases, making the identification of all minima on the potential energy surface an impossible task. Therefore, the scope of searches are often reduced by either imposing some form of constraint on the conformations generated or by biasing the search toward regions where the lowest energy conformation is more likely to be found.

Molecular Dynamic and Monte Carlo Simulations

Molecular dynamics is a computational method for simulating the motion of a system of many particles. It requires knowledge of the interaction potential from which the forces acting on each particles can be calculated, and the equations of motion that govern the dynamics of the particles. Molecular mechanics force fields are often used as the potential functions in molecular dynamics simulations. The force on atom *i* is calculated from the derivatives of the potential energy function with respect to the position of atom *i* (dE/dxi, dE/dy_i, dE/dz_i). Newton's equation, $\boldsymbol{f}_i = m_i\boldsymbol{a}_i$, is used for finding the accelerations of each particles at each simulation step. More details of the methodology of molecular dynamics and its applications in biology may be found in van Gunsteren and Berendsen (1990); Karplus and Pesko (1990); and van Gunsteren *et al*. (1994).

The total energy of a system is the sum of both potential energy and kinetic energy. The mean kinetic energy is related to the temperature *T* of the system by

$$\frac{1}{2}\sum_{i=1}^{N} m_i \left\langle v_i^2 \right\rangle = \frac{3}{2} N k_B T \quad (13)$$

where *N* is the total number of atoms in the system, $<v_i^2>$ is the average velocity squared of the *i*th atom and k_B is the **Boltzmann constant**.

Equation 13 can be used to control the temperature of the system. **Simulated annealing** is a technique where the simulated protein system starts at a high temperature, and then is cooled down gradually. By heating the protein to a high temperature, the simulation enables it to overcome larger energy barriers and to sample more conformations of interest. Ideally, as the system is cooled towards 0^{o}K, the protein is trapped in the global mininum energy conformation. If the force field used in the simulation has sufficient accuracy, this global minimum energy conformation should be close to the native structure of the protein. Metropolis and coworkers (1953) developed a **Monte Carlo method** for randomly searching the conformational space that simulates a molecular system by randomly changing its conformation. The energy of each new random conformation is compared to the energy of the previous one. If the new energy is lower, then the new structure becomes the current conformation. If the new energy is higher, then the value of the **Boltzmann factor** is compared to a random number between 0 and 1. If the Boltzmann factor is greater than the random number, then the new structure becomes the current conformation.

The advantage of a Monte Carlo method is that its randomness can overcome many energy barriers. On the other hand, for the same reason, simulations using Monte Carlo methods are usually slower to converge than those using molecular dynamics. Simulated annealing can be carried out in a Monte Carlo just as in a molecular dynamic simulation.

Predicting homologous, loop and side chain conformation

As mentioned, limits of computational power at this time preclude complete modeling of the entire protein molecule. However, if the system is contrained in some fashion, parts of it may be more easily modeled. Three constrained models of interest are those for homologous structures, for loops, and for side-chains. The discussion that follows refers to the geometry of the peptide bond, as shown in Figure 2.

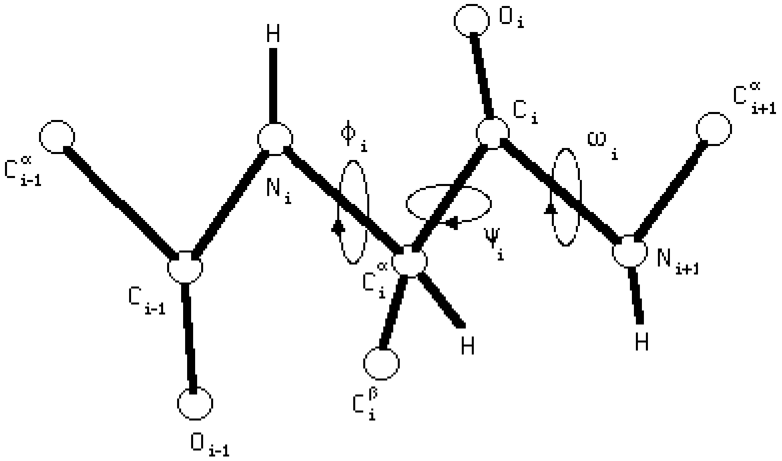

Figure 2 - A segment of polypeptide chain defining the backbone dihedral angles ϕ,ψ, and ω and a representative side-chain torsional angle χ_i .

Homologous Structures. Combining primary sequence homology ad energy calculations including solvation, Schiffer *et al.* (1990) developed a "knowledge-based" method for predicting the structure of a protein from a known structure. Once a homologous protein with known structure is chosen for a sequence, the residues in the known structure are exchanged for the sequence of the unknown protein and a computational search is carried out on the exchanged residues using molecular mechanics energy minimization. The predicted structure is the lowest energy conformer. Protein-solvent interactions are included in the energy minimization using the solvent exposed surface area of each atom and a set of experimentally derived atomic solvation energy parameters. To test their method, the structure of rat trypsin was predicted from the crystal structure of bovine trypsin. The two primary sequences are 74% identical with a difference of 56 residues between the two proteins. Volume overlap integrals (see Introduction) were used to measure the accuracy of the prediction instead of the r.m.s. After exchanging all 56 residues, the predicted rat trypsin structure has an overall 22% volume overlap error with a standard deviation of 15%. Since all of the residues exchanged between the two sequences are external residues with are highly accessible to the solvent, the authors concluded that the inclusion of solvation is crucial to the energy calculation (Schiffer *et al.*, 1990).

As menioned under Homology Modeling, Caldeira and co-workers (1994) recently used molecular modeling techniques to model one flavodoxin on another with known structure to which it had 49% sequence identity. Since the two proteins could be aligned with no gaps or insertions, the backbone coordinates were kept the same, side-chains were altered where needed, and energy was minimized in a four-step procedure using AMBER. First, main-chain and active site atoms were constrained, and energy minimization was carried out ignoring electrostatic interactions. Second, the same constraints were used but electrostatics were included. Third, most of the constraints were removed, but atoms of isoalloxazine rings were kept rigid and coplanar, ignoring the electrostatics. Finally, do as in step three, but include the electrostatics.

Loops. One of the more difficult tasks in predicting tertiary protein structure using homology modeling is to predict the conformation of loop regions. Among members of a protein family, the hydrophobic core regions with a sequence identity of 30% or more tend to adopt a similar three-dimensional structure, while surface loops in these proteins often have little or no sequence or structural similarity to each other. Yet the knowledge of the correct conformation of these loops are important in both the structural stability and biological function of a protein. Thus predicting loop conformation, a special case of homologous structure prediction, is a separate, specialized area of molecular modeling research.

Assuming that the two end points of a short chain are fixed and all peptide units are planar, Gõ and Scheraga (1970) developed a mathematical method for calculating the conformational energy of the local conformational deformations of polypeptide chain molecules. In order to reduce the cost of computation, only the backbone dihedral angles ϕ and ψ of each residues are treated as variables, keeping all bond lengths, bond angles and peptide bond dihedral angles ω fixed. The solution of the equations that generates a local deformation requires at least six degrees of freedom. Therefore, for a chain with *n* residues, *2n-6* dihedral angles (ϕ,ψ) must be specified in order to solve for the remaining six dihedral angles.

A different approach to determine loop structure in the presence of a fixed core was developed by Ducek and Scheraga (1990) using global energy minimization. The energy calculations were carried out using the ECEPP potential energy function (Sippl *et al.*, 1984). The conformational search was confined near the known distributions of ϕ and ψ angles of each amino acid. In addition, an empirically parameterized function was introduced to represent hydration free energy to increase the efficiency of hydration free energy calculation.

The global free energy minimization procedure includes the following steps:

1. A seven-residue segment is deformed to generate a large collection of backbone structures.
2. A local minimization procedure is applied to each of these structures.
3. Side-chain minimization is performed to each low-energy backbone structure from step 2.
4. Structures resulting from step 3 are locally minimized.
5. The lowest energy structure is retained as the starting point for the next cycle.

The procedure was tested using nine proteins with high-resolution tertiary structures: avian pancreatic polypeptide, crambin, trypsin inhibitor, erabutoxin B, immunoglobulin B-J fragment, ribonuclease A, lysozyme, papain D, and trysin. The results suggest that reasonably complete structure searches were achieved by the procedure (Ducek and Scheraga, 1990).

Palmer and Scheraga (1990) modified the original Gõ and Scheraga algorithm by fixing the bond lengths and bond angles at values derived from high-resolution X-ray crystallographic data for each of the 20 amino acid residues. The energy calculations were carried out using ECEPP potential energy function (Sippl *et al.*, 1984). Later Palmer and Scheraga (1992) refined this conformational search procedure for short regions of polypeptide chains. The procedure generates a series of local deformations in the polypeptide chain. After eliminating the structures having serious atomic overlaps or energetically unreasonable backbone dihedral angles, the remaining deformations are then refined by energy minimization. Finally, the r.m.s. deviations (relative to the native structures) of these energy-minimized structures are calculated. The practical advandage of this so-called rigid-geometry approximation is that it allows a large number of conformations to be sampled.

A series of five-residue chain segments were selected to test the search method. These segments include α-helices, β-sheets, β-turns, and irregular regions of the RNase A structure. In addition, the method was further tested using all the reverse turns in human lysozyme. In each test, a small number of candidates including the sturcture close to the native are efficiently generated. A good correlation between low r.m.s. deviation and low energy was observed (Palmer and Scheraga, 1992).

The Gõ and Scheraga algorithm was also adopted by Bruccoleri and Karplus (1987) in their procedure, CONGEN (CONformation GENerator), for sampling the conformational space of short polypeptide segments in proteins. Similar to Scheraga's rigid-geometry method, the main degrees of freedom for the conformational space in CONGEN are single bond torsions (ϕ, ψ) and the side-chain χ_i torsion angles. However, certain bond lengths and bond angles are allowed to vary in CONGEN while all bond lengths and bond angles are fixed in the rigid-geometry method. In order to further reduce the number of variables, the values of the three torsion angles are grouped together and the backbone conformational search is carried out by iterating over sets of energetically acceptable ω, ϕ, and ψvalues selected from **Ramachandran**-type **plots**. The conformational energy is calculated using the CHARMM potential energy function. Once a set of backbone conformations is generated, the side-chains of each backbone conformations is constructed by placing the side-chain atoms based on side-chain torsion angles, followed by energy minimization. Five molecules with known structures were chosen to test the CONGEN procedure: flavodoxin, plastocyanin, and the Fv part of immunoglobulins MC/PC 603, KOL and NEW. The procedure is capable of generating conformations where the lowest energy one matches the known structure within an r.m.s. deviation of 1 Å (Bruccoleri and Karplus, 1987).

Another loop conformation search method developed by Levinthal and co-workers (Fine *et al.*, 1986; Shenkin *et al.*, 1987) starts with a conformation generated by setting all the backbone ϕ and ψ angles to random values. These angles are next altered in a iterative fashion constrained by the distances between four atoms: the N and C_α of the N-terminal residue and the C_α and carbonyl C of the C-terminal residue. After generating a large number of initial conformations, structures exhibiting bad atomic overlaps are screened out. Combined with energy minimization and molecular dynamics, the loop generating method was applied to several complementarity determining regions of the immunoglobulin MCPC603 (Shenkin *et al.*, 1986, 1987).

The application of the Gõ and Sheraga (1970) algorithm requires a predetermined distance between the two ends of the loop and the conformational search is carried out with this distance fixed. Collura *et al.* (1993) developed a loop modeling method that starts with a competely extended loop conformation. The method combines Monte Carlo simulation with a simulated annealing algorithm. The structure of a loop is predicted from the ensemble average of the coordinates of the Monte Carlo simulation at 300^o K. Loop closure is achieved by applying a harmonic distance constraint to the backbone atoms of the terminal residues. The method was tested with loop segments from immunoglobin, bovine pancreatic trypsin inhibitor, and bovine trypsin, and has an average 1 Å r.m.s. deviation for all heavy atoms. In addition, the predicted loop structures show good hydrogen bonding compared to observation (Collura *et al.*, 1993).

Fidelis *et al.* (1994) compared molecular modeling and database searching approaches for structure prediction, testing the methods on 11 loops representing typical homologous modeling problems, including seven loop regions in dihydrofolate reductase from *Lactobacillus casei* where this enzyme significantly differs in sequence from the otherwise homologous *E. coli* enzyme, a loop with an non-proline *cis* peptide bond in β-lactamase from *S. aureus*. and three IGG hypervariable loops. Using a search methodology with a database of 57 unrelated protein structures, they conclude that the database search "results in large errors in the insert region, and is not effective for comparative modeling, even for short segments."

The molecular modeling approach of Fidelis *et al.* (1994) builds segments of chain using a set of ϕ, ψ

tortional angles (11 pairs for residues other than Pro or Gly, two for Pro, and 14 for Gly) together with rigid geometry for the peptide unit. The chain is extended out from the known "root" residues building approximately half the span from each root. All possible conformations of a residue are added to a root residue and that is continued with each added residue. At each stage, conformations that overlap severly (>1.5 Å) with the surrounding protein structure are rejected. After building half a span from each root, the N- and C-pairs that may be able to form complete segments are retained for further consideration. Each such pair is subjected to 400 steps of energy minimization including only covalent and van der Waals terms (no electrostatics) in the potential function. Root backbone atoms are constrained to be close to initial positions. Finally, segments where the linking peptide is more than 30^o from planarity, and any duplicate structures, are removed. Using this systematic conformational search, all but one of the predicted loops has an all-atom r.m.s. deviation of <1 Å compared to the actual structure. Even if only the backbone is energetically minimized, and the side-chains are built on the backbone using a library of side-chain torsional angles, an all-atom r.m.s. deviation of <1.5 Å is found for all but one loop (Fidelis *et al.*, 1994).

C_α coordinates. A third constrained model of interest is predicting side-chain orientation given the backbone or C_α coordinates. This is an important and practically valuable step, since this information is sometimes available for proteins whose complete structures are unknown. Several workers have studied this problem.

Correa (1990) first constructed a crude backbone by sequentially adding residues one at a time. Energy minimization was carried out after the placement of each additional residue until the entire backbone is built. All amino acids were treated as alanine except glycine and proline. The resulting backbone was then refined using molecular dynamics by heating the backbone to 1000^o K for 100 ps, followed by a 6 ps cooling molecular dynamics run to 0^o K with steepest descent energy minimization. Next the side chains were built up sequentially by adding atoms one level at a time. With the β-carbons already in place, the first level of side chain atoms included γ-carbons, oxygens, and sulfurs. Once these atoms were added, a 30 ps molecular dynamics run at 800^o K was performed while keeping the backbone fixed. The δ and ε atoms were then added in the same manner. Finally, the remaining atoms were added followed by a 30 ps molecular dynamics run at 1000^o K. When the procedure was tested on α lytic protease, troponin C, and flavodoxin, backbone r.m.s. deviations of less than 0.5Å and overall r.m.s. deviations of less than 1.7Å were obtained (Correa, 1990).

Holm and Sander (1991) studied the same problem from a different perspective, using the observations that the side-chain rotamers in known structures have a sharp statistical distribution ad the atoms in the protein interior are closely packed with no overlap. Given an amino acid sequence with a known C_α trace, the backbone structure is constructed by scanning a protein structure database of 34 high-resolution proteins with a total of 4,759 residues to find candidate fragments that fit the chain trace according to distance criteria and then optimally select and join these fragments into a continuous chain using a dynamic algorithm to minimize the overlaps between successive fragments. Once the backbone is constructed, side-chain coordinates are generated in two steps: (i) generate sets of plausible side-chain coordinates using a rotamer library and calculate all rotamer-rotamer interaction energies, and (ii) minimize the intramolecular energy using a Monte Carlo algorithm with simulated annealing. The procedure was tested with 17 proteins of different sizes and crystallographic resolutions. For test proteins whose X-ray structure resolutions are better than 2.5 Å, the positions of side-chain atoms in the core regions have an accuracy of 1.6 Å r.m.s. deviation and 70% of χ_1 angles are within 30^o of the X-ray structure (Holm and Sander, 1991).

Bassolino-Klimas and Bruccoleri (1992) developed an algorithm for generating the complete backbone from C_{α} coordinates using the CONGEN program and CHARMM potential energy function. When constructing the backbone from Cα coordinates, the r.m.s. deviation of partial conformations to the known C_{α} coordinates was used as a guide, so that the conformational search was directed to the areas of conformational space where the best fitting structure was more likely to be found. Six proteins of known structure of various sizes and classes were used to test the method. The three-dimensional backbone coordinates generated have r.m.s. deviations ranging from 0.30-0.87Å for the α-carbons and 0.50-0.99Å for the rest of the backbone atoms. The method was then applied to two proteins, thioredoxin and triacylglycerol acylhydrolase, whose C_{α} coordinates were the only structural information available at the time. All-atom models were proposed for the backbones of these two proteins (Bassolino-Klimas and Bruccoleri, 1992).

Yet another approach by van Gelder *et al.* (1994) starts with a crude backbone constructed by placing all intermediate backbone atoms (carbonyl C, amide N) at one-third and two-thirds of the distance between Cαi and Cαi+1. Carbonyl oxygen atoms and amide hydrogens are added at idealized bond distances with ω torsional angles of 180^{o}. All C and N atoms are then randomly shifted to avoid undefined $C\alpha_{i}$ - C-N- $C\alpha_{i+1}$ backbone dihedral angles. Energy minimization is applied to the resulting backbone chain to relieve any strain in the initial backbone, fixing all C_{α} atoms to their X-ray coordinates. Next, side chain atoms are added in extended conformations. Energy minimization with a gradually increasing non-bonded cutoff distance is performed on the resulting structure to overcome the difficulties caused by very short non-bonded interactions, followed by a long molecular dynamics run with harmonic constraints on the Cα positions at 800^{o} K to ensure the gradual formation of hydrogen bonds. After the structure was cooled to 0^{o} K, both constrained and unconstrained minimizations are performed until the structure converges. Two proteins, yeast enolase and the RNA binding domain of the A protein, were used to test the method. The constructed structures give backbone r.m.s. deviation values of 0.5-0.7Å and all-atom r.m.s. deviation values of 1.5-1.9Å (van Gelder *et al.*, 1994).

Lattice model and other reduced representations

Another way to approach the problem is to simplify the model of protein structure. One of the most widely used simplified models is a lattice representation of globular proteins. In the lattice representation, the number of degrees of freedom of a protein is reduced by representing only the α-carbons, positioning them on a fixed lattice, and replacing the side chains with entities having much smaller number of degrees of freedom. Further simplification is achieved by using mean force potentials to express the interactions between various parts of the protein and the solvent.

The applicability and potential problems of lattice models were studied by Godzik *et al.* (1993), who also compared several lattices with increasing fidelity to native protein structure. Lattice models are supported by statistical observations of the local structure of the protein backbone: (i) the distance between two consecutive α-carbons has a sharp peak around 3.8 Å; and (ii) the angle between three consecutive α-carbons has a sharp peak at 90^{o}. These two facts would favor a cubic lattice with side length 3.8 Å (Figure 3). However, (iii) the tetrahedral tortional angle between four α-carbons has a sharp peak around -130^{o} and a broader peak around $20\text{-}50^{o}$, not the 90^{o} found in a cubic lattice. Lattices other than cubic, and with side lengths shorter than 3.8 Å, are needed to model this feature.

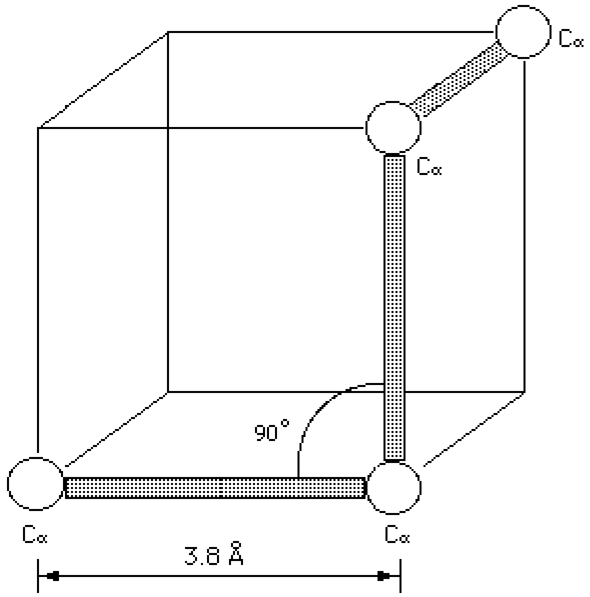

Figure 3 - Schematic drawing of a cubic lattice unit with a segment of C_{α} polypeptide chain.

Several potential problems exist for lattice models. First, the accuracy of secondary structure prediction depends heavily on the orientation of the principal axis of the lattice. Second, small errors in the model may accumulate quickly in some regions of the lattice and such errors are difficult to correct due to the inflexibility of the lattice. Third, structure fragments without regularity (i.e. turns) are often incorrectly reproduced. However, these problems can be eliminated if a lattice with sufficiently high fidelity is used (Godzik *et al.*, 1993).

Three large proteins representing three different structural classes and several smaller sequences were used to test a series of lattices: myoglobin (all-α), triose phosphate isomerase (α/β), and plastocyanin (all-β), *E. coli* repressor of primer (ROP) protein, fragments of bacteriochlorophyll A protein, myohemerythrin, and crambin. The quality of a model was measured by the r.m.s. deviation between the α-carbons in the model and those in the crystallographic structure. C_{α}r.m.s. deviations varied from 4.3 to 0.7Å for the different proteins and lattices, with two lattices able to model all sequences to deviations of 1.0Å or less. By increasing the fidelity of the lattice for backbone construction, the lattice model can be made as close to the real protein as required (Godzik *et al.*, 1993).

Using the α-carbon lattice model, a hierarchical method for simulating the protein folding and predicting the three-dimensional structure was developed (Kolinski and Skolnick, 1994a). Starting from the amino acid sequence, a coarse lattice model is first used to fold the protein of interest into several family of structures, depending on the topology and secondary structure predictions. Then the lowest energy member of each family are subject to refinement by a more precise lattice model. Subsequently, all-atom molecular dynamics folding with a detailed force field is applied to the resulting finer lattice conformations. The family with the lowest energy and smallest mean r.m.s. deviation between members is identified as the putative native structure.

The mean field potentials are derived from a statistical analysis of a database of high resolution structures. The potential function contains at least two parts: one represents the short-range interactions and the other for the long-range interactions. The short-range part describes the angular correlation between amino acid pairs down the chain. The long-range multibody potential part reflects the regular packing of the side groups. Monte Carlo dynamics is used in both coarser and finer lattice models. The dynamics of the model system is simulated by stochastically modifying the chain conformation. An energetic bias is introduced so that the system samples most frequently the valleys in an averaged Ramachandran map (Kolinski and Skolnick, 1994b).

Using sequence information alone, the hierarchical method was applied to the prediction of the three-dimensional structures of three proteins: the B domain of staphylococcal protein A, a monomeric version of the *E. coli* repressor of primer (ROP) dimer, and crambin. The predicted structures have native-like secondary structures and side chain packing. The r.m.s. deviation from the native C_α coordinates varies from 2.25 Å for protein A to 3.65 Å for ROP (Kolinski and Skolnick, 1994b).

Sun (1993) proposes a different reduced representation model for protein structure prediction using only primary sequence. In this model, each protein is reduced to its backbone atoms with bond lengths and valence angles fixed to their ideal values. Each side chain is approximated by a single virtual united atom. The backbone dihedral angles ϕ and ψ are the only coordinate variables. A statistical potential function is used to represent local and nonlocal interactions. Algorithms based on the mechanism of natural selection (Holland, 1975; Goldberg, 1989) are used for conformational searches. Tests of the model on several small proteins give native-like conformations. For the folding of melittin, a protein of 26 residues, the predicted structures have an average r.m.s. deviation of 1.66Å compared to the native x-ray crystal structure.

[TOC] [Biophysics Textbook Home Page]

Last updated November 6, 1998

Conclusions

As stated in the Introduction, our goal is prediction of tertiary structure from primary sequence. Many tools and methods presently available for protein structure prediction, but how close are we to that goal? First, some existing structures were, and are not presently, predictable in advance. For example, as mentioned earlier, we cannot yet predict the 74-residue insertion found in the thioredoxin domain of the DsbA protein (Martin *et al.*, 1993), or the cysteine knot found in gonadotrophic hormones (Lapthorn *et al.*, 1994; Wu *et al.*, 1994). Second, our present techniques were developed on sets of smaller, globular proteins, and are much more difficult to interpret when applied to, for example, a 1,200 residue sequence with multiple domains. While the estimate of no more than 500-700 unique folding topologies (Blundell and Johnson, 1993) mentioned in the Introduction may be correct, a 1,200 residue protein will have several such folding units (four or more of from 100 to 300 residues each), and, unless homologies to known structures exist, we lack reliable methods to determine domain boundaries or predict domain packing. Caballero (1992) provides a good case study of structure prediction for large proteins, using as her example the 2,703-residue *Notch* protein from *Drosophila melanogaster*. Third, proteins sometimes do not spontaneously fold to a native conformation, but require the assistance of other proteins, such as protein disulfide isomerases or chaparonins. It is not clear that predictive methods based on spontaneously-folding proteins are applicable to these more complicated cases, and it is also presently impossible to determine from primary sequence alone when such complications may occur.

At present, given an unknown primary sequence, our recommendations are: first, search for highly similar proteins in protein databases. If sequence identity of >30% is found, especially if there is also similarity in size, cellular localization, etc., similar structure is likely. If similarity is found to only a part of the unknown sequence, this may indicate a homologous domain in a multi-domain protein. If a tertiary structure is known for one or more members of this protein family, it can be expected that the new sequence will have a similar fold. Molecular modeling techniques, specifically those use for energy minimization and determining side chain and loop conformation, may be useful in detailed prediction of the tertiary structure of one member of the family based on a known structure. This recommendation is no different from that given by others in previous years (for example, Doolittle, 1987) but as sequence databases grow exponentially, with an approximate doubling time of 2 years, the chance of finding homologous proteins continues to increase as well.

If homologous sequences for all or part of the unknown exist, but their tertiary structure is presently unknown, use the homology-based secondary structure prediction method of Rost and Sander (1993a). Also use multisequence alignments of all members of the protein family to find the most highly conserved regions, and also search the PROSITE database (Bairoch, 1993) to determine important binding sites or other motifs. It may be possible to accurately predict the secondary or tertiary structure of such sites even if the entire structure remains unknown.

If no highly similar sequence is found, and no other reasons exist to expect structural similarity to less similar sequences, search PROSITE to determine if the sequence contains consensus patterns. These can indicate structure, especially if they have high sensitivity and specificity. Also, predict structural class from amino acid composition or by other techniques and use structural-class-specific secondary and tertiary structure prediction methods . Non-class-specific empirical structure prediction, together with molecular modeling, can at present only crudely predict tertiary structure without additional homology or other information. General empiric secondary structure prediction seems to have reached a maximum at 70% three-state single-residue accuracy.

Molecular modeling applied to proteins is still in its infancy. Although the results form various simplified models are encouraging, it is not clear whether the same degree of success can be achieved for entire proteins. Much work remains to be done, especially in the areas of developing better force fields and the treatment of side-chains. However, it is interesting to speculate that one might, in the not too distant future, start with a primary sequence, produce a C_α lattice model, build the side-chains and correctly position surface loops, and thus, using a combination of all the molecular modeling techniques described in this chapter, predict native tertiary structure.

Acknowledgments

Supported in part by a University of Minnesota Bush Sabbatical Award to L.E. and NIH Training Grant T15LM07041 support to K.C. We thank Len Banaszak, Harry Hogenkamp, Bob Milius, Germana Paterlini, Dan Prestridge, and David Wills for helpful discussions.

[TOC] [Biophysics Textbook Home Page]

December 4, 1998
Dianne McGavin

Glossary

back-propagation of errors algorithm
A neural network training procedure that improves the performance of a neural network by gradually changing its weights in a "backwards" manner. During the training, the output values of the network are compared with the desired values and the error information is propagated back through the network by calculating the local error one layer at a time, starting from the layer below the output layer and ending at the layer above the input layer. Weights that connect the units are updated so that the difference between the desired and the generated outputs are minimized.

Boltzmann's distribution law and constant
Boltzmann's distribution law is defined as:

$N_1 / N_2 = \exp(-\Delta\varepsilon / kT)$

where $\Delta\varepsilon = \varepsilon_2 - \varepsilon_1$; N_1 is the number of molecules in state 1 and ε1 is the energy of state 1; N_2 is the number of molecules in state 2 and ε2 is the energy of state 2; k is the **Boltzman constant** (1.38×10^{-23} J K-1); and T is the temperature in ^{o}K. The state of a protein is defined by the relative position of the atoms in the protein.

Boltzmann factor
The term $\exp(-\Delta\varepsilon / kT)$ is called the Boltzmann factor.

Chou - Fasman (CF) method
An empirical statistical method for secondary structure prediction which is based on the probability of a given amino acid residue being in a given secondary structure or random coil.

conservation (*C*)
A variable defined by Russell and Barton (1993) to measure the similarity found in a protein family, as the percentage of alignment positions sharing seven or more property states (hydrophobicity, aliphatic, etc.) as defined by Zvelebil *et al.* (1987), across all aligned sequences.

correlation coefficient (Mathews)
For the structure type *a*, the correlation coefficient is defined by

$$C_a = \frac{(p_a n_a) - (u_a o_a)}{\sqrt{(n_a + u_a)(n_a + o_a)(p_a + u_a)(p_a + o_a)}}$$

where p_a is the number of correctly predicted cases, n_a is the number of correctly rejected cases, o_a is the number of overpredicted cases, and u_a is the number of underpredicted cases.

energy minimization
The class of computational methods that starts with a set of atomic coordinates of a system and a potential energy function, and finds a nearby potential energy local minimum. Some energy minimization methods use the first-derivatives of the potential energy function for moving all the atoms towards the local minimum, others utilize both the first- and the second-derivatives of the potential function. The methods using the first-derivatives are usually less computationally intensive, while higher accuracy can often be achieved by using the methods involving both the first- and second-

derivatives.

fold
The generalized tertiary structure of a protein family or superfamily.

GOR method
The Garnier, Osguthorpe and Robson empirical statistical method of secondary structure prediction that is based on the probability of an amino acid of any type being associated with a neighbor of any type at position j, where j varies from +8 to -8 along the sequence.

helical wheel
A method of arrangement of presumed helical, sequential residues in a circle or wheel, with each residue 100 degrees from its predecessor. Wheels of amphipathic helices would show a region with a preponderance of hydrophobic residues.

homology modeling
Defined most strictly, it means predicting the tertiary structure of an unknown based on the known coordinates of a protein to which it has a high degree of sequence identity/similarity. The broader definition used here includes, in addition, predictions of secondary structure based on two or more homologous sequences, and the development of consensus sequence patterns.

molecular dynamics
A computational method for simulating the motion of a system of many particles. It requires the interaction potential from which the forces acting on each particles can be calculated, and the equations of motion that govern the dynamics of the particles. Molecular mechanics force fields are often used as the potential functions in molecular dynamics simulations. The force on atom *i* is calculated from the derivatives of the potential energy function with respect to the position of atom *i* (dE/dx_i, dE/dy_i, dE/dz_i). Newton's equation,

$\boldsymbol{f}_i = m_i\boldsymbol{a}_i$, is used for finding the accelerations of each particles at each simulation step.

molecular mechanics
A computational method designed to give accurate structures and energies of molecules. It treats molecules as collections of masses that are interacting with each other via harmonic (or more complicated) forces between bonded atoms and via van der Waals and electrostatic forces between non-bonded atoms. Mathematical functions (called potential energy functions) of the atomic coordinates are used to describe these interactions. Various parameters derived from experimental observations are included in the potential energy functions.

Monte Carlo method
A conformation search method that simulates a molecular system by randomly changing its conformation. The energy of each new conformation is compared to the energy of the previous one. If the new energy is lower, then the new structure becomes the current conformation. If the new energy is higher, then the value of the Boltzmann factor (exp[-(E_{new} - E_{old})] is compared to a random number between 0 and 1. If the Boltzmann factor is greater than the random number, then the new structure becomes the current conformation.

multisequence alignments
A method of positioning three or more primary sequences, including gaps, to optimally align regions of

highest similarity.

neural network, artificial (see also **perceptron**)
A computer program containing an input layer of units which receives input signals, an output layer of units which outputs structure predictions, and zero, one, or more hidden layers in-between the input and output layers. The units of each layer of the network are connected to each unit in the subsequent layer with a real number as a weight. A neural network is trained to produce output which recognizes patterns in input by changing the weights between various units.

perceptron
A neural network with no hidden layers. A perceptron only detects first-order correlations between input signals and output responses.

potential energy function
The equations and parameters that define the potential energy of molecules, also known as the **force field**.

primary structure
A sequential list of the amino acid residues which make up the protein, starting at the N-terminus and ending at the C-terminus. Also known as **primary sequence**.

protein folding problem, the
How primary sequence alone can determine the tertiary structure of folded proteins.

Ramachandran plots
Plots of ϕ vs ψ, which when carried out for numerous proteins can show that residues predominently fall into α-helical, β-sheet and other well-defined secondary structural classes.

root mean square (r.m.s.) deviation

$$\Delta R = \sqrt{\frac{\sum |x_i - Y_i|^2}{N}}$$

where *N* is the total number of atoms in the structure, *xi* is a set of atomic coordinates for one atom in a (possibly known) structure, and *Yi* is the set of coordinates for the corresponding atom in a second (possibly predicted) structure which has been mathematically transformed such that the sum of the squares of the distance deviations

$$\sum |x_i - Y_i|^2$$

is a minimum.

secondary structure
The folding of the primary structure in frequently-occuring forms, primarily α-helices, β-strands, and turns. Helices are sometimes subdivided into those which are mostly buried (hydrophobic) and those which have one "side" exposed to the surface (amphipathic). Residues in β-strands can be similarly divided into internal and external. Internal residues are shared by two β-ladders while external residues belong to a maximum of one β-ladder.

segment overlap (*Sov*)

$$Sov = \frac{\sum\left(\frac{\mathrm{minov}(s_1;s_2) + \delta}{\mathrm{maxov}(s_1;s_2)}\right) * \mathrm{len}(s_1)}{N}$$

where *N* is the total number of residues in the protein; the numerator is summed over all segments of secondary structure; subscripts 1 and 2 are the two sequences of secondary structures being compared (1 is usually observed and 2, predicted); s_1 and s_2 are two segments, one from each sequence, that have in common at least one residue position in the same secondary structure; min*ov* is the actual overlap between the two segments; max*ov* is the total extent of either sequence, and *len*(s_1) is the length of the observed segment. δ is an integer variable chosen to be smaller than min*ov* and smaller than 1/2 the length of s_1; $\delta = 1, 2$, or 3 for short, intermediate, and long segments. The ratio of min*ov*/max*ov* is constrained to a maximum value of 1.0.

sensitivity and **specificity**
Two statistics for assessing the accuracy of sequence patterns are **sensitivity** = TP/(TP + FN), and **specificity** = TN/(TN + FP),

where there are two sets of test sequences, one of those sequences which are known to contain the structural feature under study (knowns) and one of those sequences which are known not to contain it (controls). Then TP is the number of true positives (correct matches where a pattern is found in the knowns); TN is the number of true negatives (correct non-matches where it is not found in the controls); FP is the number of false positives (incorrect matches where it is found in the controls); and FN is the number of false negatives (incorrect non-matches where it is not found in the knowns).

simulated annealing
A molecular dynamic or other simulation that begins with the protein at a high temperature, then cools it down gradually.

single residue accuracy (see also **three-state single residue accuracy**)
The number of residues correctly predicted to contain a structure divided by the number of residues that do contain that structure

statisical methods
Methods that use statistical information on the probabilities of various amino acids being in certain structural states within a protein to develop rules for secondary structure predictions.

structural class
Usually one of four different groups of protein folds, based on the predominant secondary structure: all-α, all-β, α/β (α alternating with β), and α + β (α followed by β). For completness, a fifth small, irregular class is sometimes included.

tertiary structure
A protein's native (natural) three-dimensional structure. Sometimes tertiary fold or **fold** to refer specificially to the backbone Cα structure.

three-state single residue accuracy (Q_3)

$$Q_3 = \frac{P_\alpha + P_\beta + P_{coil}}{N}$$

where *N* is the total number of predicted residues and P_a is the number of correctly predicted secondary structures of type *a*.. Q_3 values of from 0.5 to 0.7 (50-70% accuracy) have been reported for Chou-Fasman, GOR and other current methods.

volume overlap integral
A measure of the spatial errors between two structures. The two structures are superimposed by overlapping their C_α backbones. The volume of a particular residue is calculated by extending the atomic coordinates of each atom into a sphere of radius equal to its van der Waals radius. The percentage volume overlap betwen the two residues is determined by the volume overlap between the predicted residue and the residue in the crystal structure.

[TOC] [Biophysics Textbook Home Page]

November 9, 1998

References

Introduction

Benner, S. A. (1992). Predicting *de novo* the folded structure of proteins. Curr. Opin. in Struct. Biol. 2, 402-412.

Benner, S. A. (1993). Predicting the conformation of proteins: Man versus machine. FEBS Letters 325, 29-33. MEDLINE UID = 93292645

Blundell, T.L. & Johnson, M.S. (1993) Catching a common fold. Prot. Sci. 2: 877-883. MEDLINE UID = 93306197

Chou, P. Y., & Fasman, G. D. (1974a). Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. Biochemistry 13, 211-222. MEDLINE UID = 74080189

Chou, P. Y., & Fasman, G. D. (1974b). Prediction of protein conformation. Biochemistry 13, 222-245. MEDLINE UID = 74080190

Cohen, B. I., & Cohen, F. E. (1994). Predictions of protein secondary and tertiary structure. In: Biocomputing: Informatics and Genome Projects (Smith, D.W., ed.) pp 203-232, Academic Press, New York.

Cohen, F. E., & Kunitz I. D. (1989). Tertiary structure prediction. In: Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G. D., ed.) pp 647-705, Plenum Press, New York.

Cohen, F. E., & Sternberg, M. J. E. (1980). On the prediction of protein structure: The significance of the root-mean-square deviation. J. Mol. Biol. 138, 321-333. MEDLINE UID = 81009572

Ellis, L. B. M., & Milius, R. P. (1994). Valid and invalid implementations of GOR secondary structure predictions. CABIOS, 10, 341-348. MEDLINE UID = 95007045

Fasman, G. D., ed. (1989a). Prediction of Protein Structure and the Principles of Protein Conformation, Plenum Press, New York.

Fasman, G. D. (1989b) The development of the prediction of protein structure. In: Prediction of Protein Structure and the Principles of Protein Conformation, (Fasman, G. D., ed.) pp193-307, Plenum Press, New York.

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978). Analysis of the accuracy and implication of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120, 97-120. MEDLINE UID = 78153763

Garnier, J., & Robson, B. (1989). The GOR method for predicting secondary structures in proteins. In: Prediction of Protein Structure and the Principles of Protein Conformation, (Fasman, G. D., ed.) pp 417-465, Plenum Press, New York.

Garnier, J., & Levin, J. M. (1991). The protein structure code: what is its present status? CABIOS 7, 133-142. MEDLINE UID = 91283887

Heijne, G. (1994). Membrane proteins: From sequence to structure. Annu. Rev. Biophys. Biomol. Struct. 23, 167-192.

Jenny, T. F., & Benner, S. A. (1994). Evaluating predictions of secondary structure in proteins. Biochem. Biophys. Res. Comm. 200,149-155. MEDLINE UID = 94220078

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22, 2577-2637. MEDLINE UID = 84128824

Lathrop, R. H., Webster, T., Smith, R., Winston, P., & Smith T. (1993). Integrating AI in sequence analysis. In: Artificial Intelligence and Molecular Biology (Hunter, L., ed), pp 210-258, AAAI Press, Menlo Park CA.

Lesk, A. M. (1991). Protein Architecture: A Practical Approach, pp130-131, Oxford University Press, New York.

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

Nagano, K. (1989). "Prediction of packing of secondary structure" In: Prediction of Protein Structure and the Principles of Protein Conformation, (Fasman, G. D., ed.) pp467-548, Plenum Press, New York.

Nishikawa, K., & Noguchi, T. (1991). Predicting protein secondary structure based on amino acid sequence. Methods in Enzymology 202, 31-44.

Prothero, J. W. (1966). Correlation between the distribution of amino acids and alpha helices. Biophys. J. 6, 367-370. MEDLINE UID = 67123950

Ptitsyn, O. B. (1969). Statistical analysis of the distribution of amino acid residues among helical and nonhelical regions in globular proteins. J. Mol. Biol. 42, 501-510. MEDLINE UID = 69266504

Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202: 865-884. MEDLINE UID = 89012008

Rost, B., & Sanders, C. (1993a). Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232, 584-599. MEDLINE UID = 93347244

Rost, B., & Sanders, C. (1993b). Secondary structure prediction of all-helical proteins in two states. Prot. Eng. 6, 831-836. MEDLINE UID = 94143336

Rost, B., Schneider, R., & Sander, C. (1993). Progress in protein structure prediction? TIBS 18, 120-123. MEDLINE UID = 93262667

Rost, B., Sander, C., & Schneider, R. (1994). Redefining the goals of secondary structure prediction. J. Mol. Biol. 235, 13-26. MEDLINE UID = 94118258

Russell, R. B., & Barton, G. J. (1993). The limits of protein secondary structure prediction accuracy from multiple sequence alignment. J. Mol. Biol. 234, 951-957. MEDLINE UID = 94087754

Schiffer, M., & Edmundson, A. B. (1967). Use of helical wheels to represent the structures of proteins and to idenify segments with helical propensity. Biophys. J. 7, 121-135. MEDLINE UID = 68002104

Swindells, M.B., & Thornton, J.M. (1991). Structure prediction and modeling. Current Opinions in Biotechnology 2: 512-519. MEDLINE UID = 92063201

General Empiric

Statistical

Chou, P. Y., & Fasman, G. D. (1989). Prediction of protein structural class from animo acid composition. In: Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G.D., ed.), pp. 549-586. Plenum Press, New York.

Chou, P. Y., & Fasman, G. D. (1974a). Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. Biochemistry 13, 211-222. MEDLINE UID = 74080189

Chou, P. Y., & Fasman, G. D. (1974b). Prediction of protein conformation. Biochemistry 13, 222-245. MEDLINE UID = 74080190

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D., & Fletterick, R. J. (1986). Turn prediction in proteins using a pattern-matching approach. Biochemistry 25, 266-275. MEDLINE UID = 86159695

Cohen, B. I., Presnell, S. R., & Cohen, F. E. (1991). Pattern-based approaches to protein structure prediction. Meth. Enzym. 202, 252-268.

Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J.L., Berzofsky, J. A., & DeLisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. J. Mol. Biol. 195, 659-685. MEDLINE UID = 88011275

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120, 97-120. MEDLINE UID = 78153763

Garnier, J., & Robson, B. (1989). The GOR method for predicting secondary structures in proteins. In: Prediction of Protein Structure and the Principles of Protein Conformation, (Fasman, G. D., ed.) pp 417-465, Plenum Press, New York.

Garratt, R. C., Thornton, J. M., & Taylor, W. R. (1991). An extension of secondary structure prediction towards the prediction of tertiary structure. FEBS Lett. 280, 141-146. MEDLINE UID = 91184399

Gibrat, J. F., Garnier, J. & Robson, B. (1987) Further developments of protein secondary structure prediction using information theory: New parameters and consideration of residue pairs. J. Mol. Biol. 198, 425-443. MEDLINE UID = 88118942

Leng, B., Buchanan, B.G., & Nicholas, H.B. (1994). Protein secondary structure prediction using two-level case-based reasoning. J. Comp. Biology, 1, 25-38. MEDLINE UID = 96382587

Levin, J. M., & Garnier, J. (1988). Improvements in a secondary structure prediction method based on a search for local sequence homologies and its model building tool. Biochim Biophys. Acta. 955, 283-295. MEDLINE UID = 88294096

Kneller, D.,G., Cohen, F. E., Langridge R. (1990). Improvements in protein secondary structure predicted by an enhanced neural network. J. Mol.Biol. 214, 171-182. MEDLINE UID = 90317819

Presnell, S. R., Cohen, B. I., & Cohen, F. E. (1992). A segment-based approach to protein secondary structure prediction. Biochemistry 31, 983-993. MEDLINE UID = 92135218

Prevelige, P., & Fasman, G. D. (1989). Chou-Fasman prediction of the secondary structure of proteins: The Chou-Fasman-Prevelige algorithm. In: Prediction of Protein Structure and the Principles of Protein Conformation (Fasman,G.D., ed.), pp 391-416, Plenum Press, New York.

Ralph, W. W., Webster, T., & Smith,T. F. (1987). A modified Chou and Fasman protein structure algorithm. Comput. Appl. Biosci. 3, 211-216. MEDLINE UID = 88270157

Schiffer, M., & Edmunson, A.B., (1967). Use of helical wheels to represent the structures of proteins and to identify segments with helical propensity Biophys. J. 7, 121-135.

Neural Networks

Eisenberg, D., Weissk, R. M., Terwilliger, T. C., & Wilcox, W. (1982). Hydrophobic moments and protein structure. Faraday Symp. Chem. Soc. 17, 109-120.

Holley, L. H., & Karplus, M. (1989). Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. USA. 86, 152-156. MEDLINE UID = 89098870

Hirst, J.D., and Stenberg, M.J. (1991) Prediction of ATP-binding motifs: A comparison of a perceptron-type neural network and a consensus sequence method. Protein Engineering, 6, 615-623.

Holley, L. H., & Karplus, M. (1989). Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. USA. 86, 152-156. MEDLINE UID = 89098870

Kneller, D. G., Cohen, F. E., & Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. 214, 171-182. MEDLINE UID = 90317819

Muggleton, S., King, R.D., & Sternberg, M.J. (1992). Protein secondary structure prediction using logic-based machine learning. PRot. Eng. 5, 647-657. MEDLINE UID = 93126314

Muggleton, S., King, R.D., & Sternberg, M.J. (1992). Correction to: Protein secondary structure prediction using logic-based machine learning. PRot. Eng. 6, 549.

Presnell, S.R., Cohen, B.I., & Cohen, F.E. (1993). MacMatch: a tool for pattern-based protein secondary structure prediction. Comp. Appl. Biosci. 9, 373-4. MEDLINE UID = 93313708

Qian, N., & Sejnowski, T.J. (1988). Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202, 865-884. MEDLINE UID = 89012008

Rost, B., & Sander, C. (1993a) Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232, 584-599.

Rost, B., & Sander, C. (1993b) Secondary structure prediction of all-helical proteins in two states. Prot. Eng. 6, 831-836.

Wilcox, G., Pliac, M. & Liebman, M.N. (1990). Neural network analysis of protein tertiary structure. Tetrahedron Computer Methodology, 3, 191-211.

Vieth, M., & Kolinski, A. (1991). Prediction of protein secondary structure by an enhanced neural network. Acta Biochim Pol. 38, 335-351. MEDLINE UID = 92188670

Vieth, M., Kolinski, A., Skolnick, J. Sikorski, A. (1992). Prediction of protein secondary structure by neural networks: Encoding short and long range patterns of amino acid packing. Acta Biochim Pol. 39, 369-392. MEDLINE UID = 93190678

Tertiary Structure

Chou, P. Y. (1989). Prediction of protein structural class from amino acid composition. In: Prediction of Protein Structure and the Principles of Protein Conformation (Fasman,G.D., ed.), pp 549-586, Plenum Press, New York.

Metfessel, B. A., Saurugger, P. N., Connelly, D. P. & Rich, S. S. (1993). Cross-validation of protein structural class prediction using statistical clustering and neural networks. Protein Science 2, 1171-1182. MEDLINE UID = 93364270

Muskal, S. M., & Kim, S-H. (1992). Predicting protein secondary structure content. J. Mol. Biol. 225, 713-727. MEDLINE UID = 92292156

Vieth, M., Kolinski, A., Skolnick, J., & Sikorski, A. (1992). Prediction of protein secondary structure by neural networks: Encoding short and long range patterns of amino acid packing. Acta Biochim. Pol. 39, 369-392.

Wilcox. G., Pliac, M., & Liebman, M.N. (1990). Neural network analysis of protein tertiary structure. Tetrahedron Computer Methodology 3, 191-211.

Xin, Y., Carmeli, T.T., Liebman, M.N., & Wilcox, G.L. (1993). Use of the backpropagation nerual network algorithm for predicton of protein folding patterns. In: Proceedings of the Second International Conference on Bioinformatics, Supercomputing, and Complex Genome analysis (Lim, H.A., Fickett, J.W., Cantor, C.R., & Robbins, R.J., eds.), pp. 391-416. World Scientific Publishing, Inc., Singapore.

Zhang, C.-T., & Chou, K.-C. (1992) An optimization approach to predicting protein structural class from amino acid composition. Protein Science 1, 401-408. MEDLINE UID = 93278274

Homology Modeling

Bairoch, A. (1993). The PROSITE dictionary of sites and patterns in proteins, its current status. Nucl. Acids Res. 21, 3097-3103. MEDLINE UID = 93324415

Barton, G.J., Cohen, P.T.W., & Bradford, D. (1994). Conservation analysis and structure prediction of the protein serine/threonine phosphatases. Eur. J. Biochem. 220, 225-237.

Brenner, S.A. (1992). Prediction de novo the folded structure of proteins. Curr. Opin. Struct. Biol. 2, 402-412.

Benner, S. A., & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. Adv. Enzyme Regul. 31,121-81. MEDLINE UID = 91344683

Benner, S.A., Badcoe, I, Cohen, M. A., Gerloff, D. L. (1994). Bona fide prediction of aspects of protein conformation: Assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. J. Mol. Biol. 235, 926-58. MEDLINE UID = 94118362

Benner, S. A., Cohen, M. A., & Gerloff, D. (1993). Predicted secondary structure for the Src homology 3 domain. J. Mol. Biol. 229, 295-305. MEDLINE UID = 93156043

Boldt, Y.R., Sadowsky, M.J., Ellis, L.B.M., Que, L., & Wackett, L.P. (1995). A manganese-dependent dioxygenase from *Arthrobacter globiforuis* CM-2 belongs to the major extradiol dioxygenase family . J. Bacteriology 177, 1225-1232.

Boniface, J.J., & Reichert, L.E. (1990). Evidence for a novel thioredoxin-like cataltic property of ganadotropic hormone. Science 247, 61-64.

Boscott, P. E., Barton, G. J., & Richards, W. G. (1993). Secondary structure prediction for modeling by homology. Protein Eng. 6, 261-266. MEDLINE UID = 93281561

Burbaum, J.J., Starzyk, R.M., & Schimmel, P. (1990). Understanding structural relationships in proteins of unsolved three-dimensional structure. Proteins: Structure, Function, Genetics 7, 99-111.

Calderia, J., Palma, P.N., Regalla, M., Lampreia, J., Calvete J., Schafer, W., Legall, J., Moura, I., & Moura, J.J.G. (1994). Primary sequence, oxidation-reduction potentials and tertiary structure prediction of *Desulfovibrio desulfuricans* ATCC 27774 flavodoxin. Eur. J. Biochem. 220, 987-995.

Cohen, B. I., Presnell, S. R., & Cohen, F. E. (1991). Pattern-based approaches to protein structure prediction. Meth. Enzym. 202, 252-268.

Cohen, B. I., Presnell, S. R., & Cohen, F. E. (1993). Origins of structural diversity within sequentially identical hexapeptides. Protein Science 2, 2134-2145. MEDLINE UID = 94129396

Cohen, B. I., & Cohen, F. E. (1994). Predictions of protein secondary and tertiary structure. In: Biocomputing: Informatics and Genome Projects (Smith, D.W., ed.) pp 203-232, Academic Press, New York.

Crawford, I.P., Niermann, T., & Kirschner, K. (1987). Prediction of secondary structure by evolutionary comparison: Application to the a subunit of tryptophan synthase. Proteins Structure, Function genetics 2, 118-129.

Donnelly, D., Overington, J.P., & Blundell, T.L. (1994). The prediction and orientation of a-helices from sequence alignment: The combined use of environment-dependent substitution tables, Fourier transform methods and helix caping rules. Protein Engineering 7, 645-653.

Doolittle, R.F. (1987). Of URFs and ORFs. pp 35-36, University Science Books, Mill Valley, CA.

Ellis, L. B. M., Saurugger, P., & Woodward, C. K. (1992). Identification of the three-dimensional thioredoxin motif: Related structure in the ORF3 protein of the *Staph. aureus mer* operon. Biochemistry 31, 4882-4891. MEDLINE UID = 92273602

Eklund, H., Gleason, F. K., & Holmgren, A. (1991). Structural and functional relations among thioredoxins of different species. Proteins 11, 13-28. MEDLINE UID = 92073284

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120, 97-120. MEDLINE UID = 78153763

Garnier,J., & Robson,B. (1989). The GOR Method for Predicting Secondary Structure in Proteins. In: Prediction of Protein Structure and the Principles of Protein Conformation (Fasman,G.D., ed.), Plenum Press, New York.

Gerloff, D. L., Jenny, T. F., Knecht, L. J., Gonnet, G. H., & Benner, S. A. (1993a). The nitrogenase MoFe protein: A secondary structure prediction. FEBS Lett. 318, 118-24. MEDLINE UID = 93178610

Gerloff, D. L., Jenny, T. F., Knecht, L. J., & Benner, S. A. (1993b). A secondary structure prediction of the hemorrhagic metalloprotease family. Biochem Biophys Res Commun. 194, 560-5. MEDLINE UID = 93326174

Hilbert, M., Bohm, G., & Jaenicke, R. (1993). Structural relationships of homologous proteins as a fundamental principle in homology modeling. Proteins - Structure Function and Genetics, 17, 138-151. MEDLINE UID = 94089640

Hodgkin, E.E., Gillman, I.C., & Gilbert, R.J. (1994). Retrospective analysis of a secondary structure prediction: The catalytic domain of matrix metalloproteinases. Protein science 3, 984-986.

Jenny, T. F., & Benner, S.A. (1994b). A prediction of the secondary structure of the pleckstrin homology domain. Proteins-Structure Function and Genetics 20, 1-3. MEDLINE UID = 95124982

Jentoft, J.E., Shoham, M., Hurst, D., & Patel, M.S. (1992). A structural model for human dihydrolipoamide dehydrogenase. Proteins: Structure Function Genetics 14, 88-101.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers **22**: 2577-2637. MEDLINE UID = 84128824

Lapthorn, A. J., Harris, D.C., Littlejohn, A., Lustbader, J. W., Canfield, R. E., Machin, K. J., Morgan, F.J., & Isaacs, N. W. (1994). Crystal structure of human chorionic gonadotropin. Nature 369, 455-461. MEDLINE UID = 94261179

Lathrop, R. H., Webster, T. A., & Smith, T. F. (1987). ARIADNE: Pattern-directed inference and hierarchical abstraction in protein structure recognition. Commun. Assoc. Comput. Machinery 30, 909-921.

Lathrop, R. H., Webster, T., Smith, R., Winston, P., & Smith, T. (1993) Integrating AI in sequence analysis. In: Artificial Intelligence and Molecular Biology, (Hunter, L., ed.) pp 210-258, AAAI Press, Menlo Park CA.

Levin, J. M., & Garnier, J. (1988). Improvements in a secondary structure prediction method based on a search for local sequence homologies and its model building tool. Biochim Biophys. Acta. 955, 283-295. MEDLINE UID = 88294096

Levin, J. M., Pascarella, S., Argos, P., & Garnier, J. (1993). Quantification of secondary structure prediction improvement using multiple alignments. Protein Eng. 6: 849-854. MEDLINE UID = 94143338

Martin, J.L., Bardwell, J. C. A., & Kuriyan, J. (1993). Crystal structure of the DsbA protein required for disulphide bond formation *in vivo*. Nature 365, 464-468. MEDLINE UID = 94019776

Muhelbach, S.M., Gross, M., Wirz, T., Wallimann, T., Perriard, J.-C., & Wyss M. (1994). Sequence homology and structure predictions of the creatine kinase isoenzymes. Molec. Cell. Biochem. 133/134, 245-262.

Niermann, T., & Kirschner, K. (1991). Improving the prediction of secondary structure of 'TIM-barrel' enzymes. Prot. Eng. 4, 359-370. MEDLINE UID = 91312884

Pickett, S.D., Saqi, M.A.S., & Sternberg, M. J. E. (1992). Evaluation of the sequence template method for protein structure prediction: Discrimination of the (β/α)8-barrel fold. J. Mol. Biol. 228, 170-187. MEDLINE UID = 93078255

Rees, D.C. (1990). Three-dimensional protein structure prediction workshop: Overview and summary. In: Current Researsch in Protein Chemistry (Villafranca, J., ed.), p.555. Academic Press, New York.

Smith, R. F., & Smith, T. F. (1992). Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modeling. Prot. Eng. 5: 35-41. MEDLINE UID = 92335156

Thornton, J.M., Flores, T.P., Jones, D.T., & Swindells, M.B. (1991). Prediction of protress at last. Nature 354, 105-106.

Wilson, I. A., Haft, D. H., Getzoff, E. D., Tainer, J. A., Lerner, R. A., & Brenner, S. (1985). Identical short peptide sequences in unrelated proteins can have different conformations: A testing ground for theories of immune recognition. Proc. Natl. Acad. Sci. USA 82, 5255-5259. MEDLINE UID = 85270503

Wu, H., Lustbader, J.W., Lie Y., Canfield, R.E., & Hendrickson, W.A. (1994). Structure of human chorionic gonadotropin at 2.6Å resolution from MAD analysis of the selenomethionyl protein. Structure 2, 545-558. MEDLINE UID = 95006321

Zvelebil, M. J., Barton, G. J., Taylor, W. R., Sternberg, M. J. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J.Mol.Biol. 195, 957-961. MEDLINE UID = 88011304.

Molecular Modeling

Andrews, D. H. (1930). The relation between the Raman spectra and the structure of organic molecules. Phys. Rev. 36, 544-554.

Bassolino-Klimas, D., & Bruccoleri, R. E. (1992). Application of a directed conformational search for generating 3-D coordinates for protein structures from α-carbon coordinates. Proteins - Structure Function and Genetics 14, 465-474. MEDLINE UID = 93066163

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comp. Chem. 4, 187-217.

Bruccoleri, R. E., & Karplus, M. (1987). Prediction of folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26, 137-168. MEDLINE UID = 87101371

Burkert, U., & Allinger, N.L. (1982). Molecular Mechanics. ACS Monograph 177, American Chemical Society, Washington, DC.

Calderia, J., Palma, P.N., Regalla, M., Lampreia, J., Calvete J., Schafer, W., Legall, J., Moura, I., & Moura, J.J.G. (1994). Primary sequence, oxidation-reduction potentials and tertiary structure prediction of *Desulfovibrio desulfuricans* ATCC 27774 flavodoxin. Eur. J. Biochem. 220, 987-995.

Cohen, B. I., Presnell, S. R., & Cohen, F. E. (1991). Pattern-based approaches to protein structure prediction. Meth. Enzym. 202, 252-268.

Collura, V., Higo, J., & Garnier, J. (1993). Modeling of protein loops by simulated annealing. Protein Science. 2, 1502-1510.

Correa, P.E. (1990). The building of protein structures from α-carbon coordinates. Proteins - Structure Function and Genetics 7, 366-377. MEDLINE UID = 90341219

Ferguson, D. M., & Raber, D. J. (1989). A new approach to probing conformational space with molecular mechanics: Random incremental pulse search. J. Am. Chem. Soc. 111, 4371-4378.

Fidelis, K., Stern, P.S., Bacon, D., & Moult., J. (1994). Comparison of systematic search and database methods for constructing segment of protein structure. Prot.Eng. 7, 953-960.

Fine, R.M., Wang, H., Shenkin, P.S., Yarmush, D.L., & Levinthal, C. (1986). Predicting antibody hypervariable loop conformations II: Minimization and molecular dynamics studies of MCP603 from many randomly generated loop conformations. Proteins. 1, 342-362.

Go, N., & Scheraga, H. A. (1970). Ring closure and local conformational deformations of chain molecules. Macromolecules 3, 178-187.

Godzik, A., Kolinski, A., & Skolnick, J. (1993). Lattice representations of globular proteins: How good are they? J. Comp. Chem. 14, 1194-1202.

Goldberg, D. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, New York.

Holland, J. H. (1975). Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor.

Holm, L., & Sander, C., (1991). Database algorithm for generating protein backbone and side-chain coordinates from a C-alpha trace: Application to model building and detection of coordinate errors. J. Mol. Biol. 218, 183-194.

Karplus, M., & Petsko, G. A. (1990). Molecular dynamics simulations in biology. Nature, 347, 631-639. MEDLINE UID = 91015381

Kolinski, A., & Skolnick, J. (1994a). Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins - Structure Function and Genetics 18, 338-352. MEDLINE UID = 94269051

Kolinski, A., & Skolnick, J. (1994b). Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. Proteins - Structure Function and Genetics 18, 353-366. MEDLINE UID = 94269052

Kollman, P.A. (1991). AMBER 4.0, University of California Press, San Francisco.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., & Teller, A. H. (1953). Equation of state calculations by fast computing machines. J. Chem. Phys. 21, 1087-1092.

Nemethy, G., Gibson, K.D., almer, K.A., Yoon, C.N., Paterlini, G., Zagari, A., Rumsey, S., & Scheraga, H.A. (1992). Energy parameters in polypeptides, 10: Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm with a pplication to proline-containing peptides. J. Phys. Chem. 96, 672-6484.

Palmer, K. A., & Scheraga, H. A. (1990). Standard-geometry chains fitted to X-ray derived structures: Validation of the rigid-geometry approximation: I. Chain closure through a limited search of "loop" conformations. J. Comp. Chem. 12, 505-526.

Palmer, K.A., & Scheraga, H.A. (1992). Standard-geometry chains fitted to X-ray derived structures: Validation of the rigid-geometry approximation: II. Systematic searches for short loops in proteins: Applications to bovine pancreatic ribonuclease A and human lysozyme. J. Comp. Chem. 13, 329-350.

Schiffer, C.A., Caldwell, J.W., Kollman, P.A., & Stroud, R.M. (1990). Prediction of homologous protein structures based on conformational searches and energetics. Proteins 8, 30-43.

Shenkin, P.S., Yarmush, D.L.. Fine, R.M., Wang, H., & Levinthal, C. (1987). Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers 26, 2053-2085.

Sippl, M. J., Nemethy, G., & Scheraga, H. A. (1984). Intermolecular potentials from crystal data: 6. Determination of empirical potentials for O-H . . . O=C hydrogen bonds from packing configurations. J. Phys. Chem. 88, 6231-6233.

Sippl, M. J. (1993). Boltzmann's principle, knowledge-based mean fields and protein folding. An aproach to the computational determination of protein structures. J. Comp. Aided Mol. Design 7, 473-501.

Sun, S. (1993). Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. Protein Science 2, 762-785. MEDLINE UID = 93264948 van Gelder, C. W. G., Leusen, F. J. J., Leunissen, J. A. M., & Noordik, J. H. (1994). A molecular dynamics approach for the generation of complete protein structures from limited coordinate data. Proteins - Structure Function and Genetics 18, 174-185. MEDLINE UID = 94211758

van Gunsteren, W. F., & Berendsen, H. J. C. (1990). Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. Angew. Chem. Int. Ed. Engl. 29, 992-1023.

van Gunsteren, W. F., Luque, F. J., Timms, D., & Torda, A.E. (1994), Molecular mechanics in biology - from structure to function, taking account of solvation. Ann. Rev. Biophy. Biomol. Struc. 23, 847-863. MEDLINE UID = 95003503

Conclusions

Bairoch, A. (1993). The PROSITE dictionary of sites and patterns in proteins, its current status. Nucl. Acids Res. 21, 3097-3103. MEDLINE UID = 93324415

Blundell, T.L. & Johnson, M.S. (1993) Catching a common fold. Prot. Sci. 2: 877-883. MEDLINE UID = 93306197

Caballero, L. (1992). Practical aspects: Analysis of Notch. In: Sequence Analysis Primer (Gribskov, M. & Devereux, J. eds.), pp. 159-203. W.H. Freeman, New York.

Doolittle, R.F. (1987). Of URFs and ORFs. pp 35-36, University Science Books, Mill Valley, CA.

Lapthorn, A. J., Harris, D.C., Littlejohn, A., Lustbader, J. W., Canfield, R. E., Machin, K. J., Morgan, F.J., & Isaacs, N. W. (1994). Crystal structure of human chorionic gonadotropin. Nature 369, 455-461. MEDLINE UID = 94261179

Martin, J.L., Bardwell, J. C. A., & Kuriyan, J. (1993). Crystal structure of the DsbA protein required for disulphide bond formation *in vivo*. Nature 365, 464-468. MEDLINE UID = 94019776

Rost, B., & Sanders, C. (1993a). Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232, 584-599. MEDLINE UID = 93347244

Wu, H., Lustbader, J.W., Lie Y., Canfield, R.E., & Hendrickson, W.A. (1994). Structure of human chorionic gonadotropin at 2.6Å resolution from MAD analysis of the selenomethionyl protein. Structure 2, 545-558. MEDLINE UID = 95006321

November 12, 1998