Protein Tertiary Structures: Prediction from Amino Acid Sequences

Secondary article

Article Contents

Hongyu Zhang, *Celera Genomics, Rockville, Maryland, USA*

Protein tertiary structures contain key information for the understanding of the relationship between protein amino acid sequences and their biological functions. A large collection of computational algorithms has been developed to predict protein tertiary structures from their sequences in computers.

Introduction

Proteins are polypeptide chains consisting of a large number of amino acid residues that are covalently linked together via amide bonds. The order in which the 20 different amino acids are arranged in a protein chain is also called the primary structure of the protein. The polypeptide backbones of proteins exist in particular conformations known as the secondary structures. The secondary structures as well as their side-chains are then packed into three-dimensional structures referred to as the tertiary structures.

The biological function of a protein is often intimately dependent upon its tertiary structure. X-ray crystallography and nuclear magnetic resonance are the two most mature experimental methods used to provide detailed information about protein structures. However, to date the majority of the proteins still do not have experimentally determined structures available. As at December 2000, there were about 14 000 structures available in the protein data bank (PDB, http://www.pdb.org), and there are about 10 106 000 sequence records sequences in GenBank (http://www.ncbi.nlm.nih.gov/Genbank). Thus theoretical methods are very important tools to help biologists obtain protein structure information. The goal of theoretical research is not only to predict the structures of proteins but also to understand how protein molecules fold into the native structures.

The current methods for protein structure prediction can be roughly divided into three major categories: comparative modelling; threading; and *ab initio* prediction. For a given target protein with unknown structure, the general procedure for predicting its structure is described in **Figure 1**.

Comparative Modelling

From the available experimental data it has been observed that proteins with similar amino acid sequences usually adopt similar structures. Therefore, the easiest and also the most accurate way to predict the protein tertiary structure is to build the structure based on sequence relatives that have high sequence similarities to the target protein according to the sequence alignment results. Such an approach is called comparative modelling. In most cases those sequence relatives and the target protein belong to the same functional family in biology, i.e. they are homologues of each other. Thus, traditionally, comparative modelling is also called homology modelling.

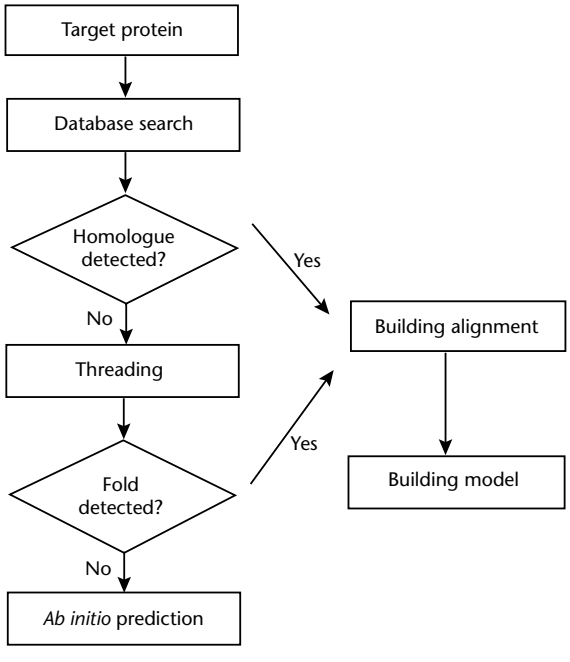

Figure 1 Procedure for predicting a protein structure from its amino acid sequence.

Database search

An initial step for comparative modelling is to check whether there is any protein in the current PDB having the similar sequence or function to the target protein. A protein found will then serve as the structural template for modelling the target protein. In most situations, the searching of the template has to proceed using a sequence comparison algorithm that is able to identify the global sequence similarity. In some cases, even when there is no global sequence similarity between two protein sequences, a close match between some important sequence fragments or local sequence patterns (also called motifs) is still significant enough for us to identify the homologous relationship between protein sequences.

To start a database search, one first needs a score function that can evaluate the similarity between amino acids. Various score functions are available. The simplest one is the identity score function, which gives score 1 for an amino acid matched to the same type of amino acid and score 0 for an amino acid mutated to a different type of amino acid. More advanced score functions are based on the statistics of the amino acid substitution frequencies in known aligned homologous sequence families. Among them, most popular ones are Dayhoff (Dayhoff *et al.*, 1978) and Blosum (Henikoff and Henikoff, 1992) matrices. The 20×20 elements in the matrices represent the substitution scores between 20 natural amino acids.

To search a large sequence database, the computer algorithms have to be able to find the close sequences correctly and quickly. Some quite efficient algorithms have been developed to solve the database search time problem, such as BLAST (Altschul *et al.*, 1990, 1997) and FASTA (Pearson and Lipman, 1988). BLAST is currently the most popular database search protocol. Its central idea is to transform the whole sequence comparison problem into an easier problem of local fragment matching and extension. FASTA achieves much of its speed and selectivity by using a lookup table to locate all identities or groups of identities between two sequences (Pearson, 1990).

Sequence alignment

After finding the template sequences for the target sequence in the structure database, the second step in comparative modelling is to align the target sequence to the template sequence. An alignment algorithm is used to find an optimal alignment for the two sequences. The result will indicate the matching, insertion or deletion of the amino acids between the target sequence and the template sequence. Thus, from a sequence alignment one can decide the structural features of each amino acid in the target protein based on the structural features of its corresponding template residue. If there are multiple templates, a multiple sequence alignment can further improve the accuracy of sequence–structure alignment.

There is no trivial solution for aligning two protein sequences because of the vast number of combinations between amino acid pairs. Fortunately, a classical algorithm, originally from the computer science field, called the dynamic programming algorithm can guarantee to quickly find the optimal alignment given a score function (Needleman and Wunsch, 1970; Smith and Waterman, 1981). The basic philosophy of the algorithm is to build up an optimal alignment using previous solutions to smaller subsequences.

The central step in Needleman–Wunsch algorithm is the construction of a score matrix. Each element in the score matrix, $F(i,j)$, is the score of the best alignment between the initial segment $x_{1,\ldots i}$ of sequence x and $y_{1,\ldots j}$ of sequence y. The 'trick' of the algorithm is that $F(i,j)$ can be built recursively according to eqn [1].

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - \delta \\ F(i,j-1) - \delta \end{cases} \quad [1]$$

In eqn [1], $s(x_i,y_j)$ is the score of aligning a residue pair (x_i,y_j), and δ is the score of a residue aligned to a gap. The principle is that the best alignment score $F(i,j)$ can only come from the three possible ways shown in the above equation: either the last residues of the two sequences (x_i and y_j), aligned together, or any of them aligned to a gap. At the beginning, $F(0,0)$ is initialized to 0, and $F(i,0)$, $F(0,j)$ are initialized to $-i\delta$ and $-j\delta$ because they represent i or j residues that are aligned to gaps.

The Needleman–Wunsch algorithm is used to look for the best match between two sequences from one end to the other. A more common situation is looking for the best alignment between the subsequences of two sequences, which can locate the common regions shared between two proteins that could have little global similarity. A very similar dynamic algorithm called the Smith–Waterman algorithm was developed to solve such local alignment problems (Smith and Waterman, 1981).

These algorithms have become the standard algorithms in this field after 20 years of improvement. Researchers can download their mature implementation programs from the Internet, such as the popular CLUSTAL program (http://www-igbmc.u-strasbg.fr/BioInfo/ClustalW/) (Higgins *et al.*, 1989).

After automatically constructing the initial alignment using the dynamic programming technique, some human intervention is helpful to adjust the errors in the computer-generated alignment; graphic tools with the addition of human expertise can identify some possibly inappropriate matches, such as a hydrophobic residue in the target being matched to the surface region in the template structure.

Building the homology model

Once the alignment is completed, one can start to build the structure model for the target protein based on the template structure. The major steps in building the homology model are conserved region modelling and loop region modelling. Conserved regions refer to those regions with conserved amino acids in the sequence alignment, most often those regions having standard secondary structures (α helix and β strands) in the template structure. Those regions will very probably keep their conformations unchanged from the template structure to the target structure, and are therefore easy to build at the very beginning. It is usually straightforward to copy the structure in those regions from the template to the target.

Loop regions are hard to model because they are less conserved in structure. In most situations they are located on the protein surface exposed to the solvent and do not have standard secondary structures. Traditionally, loop-modelling methods were categorized into two kinds of approaches: knowledge-based approach and *ab initio* approaches (for details, see the review section in Zhang *et al*., 1997). The knowledge-based approach extracts the knowledge from the current protein structure database and then applies it in the building of the new loops; the *ab initio* approach usually uses some kinds of theoretical conformational search method such as the Monte Carlo or simulated annealing methods (Leach, 1996) to build up the new loops. *Ab initio* methods are more general methods because they are not prohibited by the current size of the structure database, but traditionally they are much slower than the knowledge-based methods and therefore are not suitable for modelling very long loops. Some improved *ab initio* algorithms have achieved very high efficiency and can successfully model long protein loops very quickly (Zhang *et al*., 1997). Knowledge-based and *ab initio* algorithms can be combined together to improve the modelling accuracy; for example, one can apply both methods in the same loop region and, if they produce the similar result, have higher confidence to one's predictions.

After constructing the structures in both conserved regions and loop regions, the last steps of comparative modelling include side-chain modelling and model evaluation/refinement. Methods for side-chain modelling include Monte Carlo, genetic algorithm, side-chain rotamer library and others (Leach, 1996). They have already reached very high precision (Dunbrack, 1999). The model evaluation usually includes the checking of Ramanchandra graphs and atomic packing. Molecular mechanics and molecular dynamics are common tools (Leach, 1996) for refining the final model.

It should be pointed out that the above procedure is not a simple one-way street; in most cases it is an iterative procedure. For example, one can start with an initial sequence alignment and build the structure model; after evaluating the structure model one can go back to correct the misaligned residues or inappropriately generated side-chains and repeat the modelling procedure again.

For some years there have been very good commercial packages available in this field that bundle the comparative modelling modules into one piece of software, plus some other extra functions. They often run on powerful Unix workstations and provide very user-friendly graphic interfaces. Among the most popular ones are QUANTA and Insight-II produced by MSI and Sybyl produced by Tripos.

Threading

Of all the proteins in the current sequence database, only about 10–20% of sequences can be modelled by comparative modelling methods. For all the other sequences, it is difficult to find sequence relatives using plain sequence comparison methods.

Threading improves the sequence alignment sensitivity by introducing structural information into the alignment, where the structural information refers to the secondary or tertiary structural features of proteins. This helps because amino acids have different propensities for different secondary structures or tertiary structure environments. For example, some amino acids are more often observed in α helices than in other secondary structure units, while some amino acids appear more frequently in hydrophobic environments than do others.

The threading method is sometimes called the fold recognition method. Its basic assumption is that the number of protein folds existing in nature is limited, from several hundreds to over 1000, according to different theories (Wang, 1998). The goal of fold recognition is to identify the correct fold for the target sequence.

Most of the threading algorithms are based on the dynamic algorithm, but the key difference is the scoring strategy: in most threading algorithms the score functions include the structure information in addition to the sequence information. The earliest threading approach is the '3D profiles' method (Bowie *et al*., 1991; Luthy *et al*., 1992), in which the structural environment in each residue position of the template is classified into 18 classes based on the position's burial status, local secondary structure and polarity. The threading score matrix is then deduced from the probability of all amino acids present in those 18 classes of structure environment. For example, if a hydrophobic residue is aligned to a buried template position, the score matrix is supposed to give a high score to encourage such a type of sequence–structure match. The threading methods of Jones *et al*. (1992) and Godzik *et al*. (1992) are based on the protein residue pairwise interaction energy methods such as the potential of mean force method of Sippl (1990). The energy formulae are derived from statistical analyses of current protein structure database and reflect the

residue–residue distance distribution probabilities in known protein structures. In each step of the threading procedure, the alignment score is calculated by adding up all the pairwise interaction energies between each target residue and the template residues surrounding them.

In addition to the above methods using the sequence–structure match scores, some other threading methods also use the structure–structure match scores to evaluate the alignment between the target and the template. In those methods, although the target structure is unknown, one can still characterize it using some predicted structure properties, such as the predicted secondary structures or the predicted residue burial status (Rost and Sander, 1994).

Another important threading method is the Profile Hidden Markov Model method (HMM, see review of Durbin *et al*., 1998). This is a very sensitive tool in searching for remote homologues because of its strong statistics background. A HMM is basically a probability distribution model. To build the profile HMM, first all the sequences in the database need to be clustered into a handful of families. Each family is then used to train a HMM. Finally, the target sequence is aligned to those HMMs to identify the family to which it belongs. Although the structural information usually is not explicitly characterized in HMMs, it is implied in the corresponding statistical models. A HMM algorithm developed by Di Francesco *et al*. (1997a,b) used the structure information directly, in which the target structure is characterized by the predicted secondary structure while the template structures are represented by profile HMMs trained on the template's secondary structure patterns.

Some advanced sequence search methods such as PSI-BLAST (Altschul, 1997) utilize more sensitive position-dependent score matrices, which are very good at detecting remote homologues. Some people also consider them to belong to threading methods because of their high searching sensitivity compared to basic database searching algorithms.

Although threading methods are good at detecting remote homologues, they are often not able to give good sequence–structure alignment. The main reason is that the structure information is included in threading with many approximations, and thus can introduce significant noise into the final alignment. For example, most threading methods use the so-called 'frozen' approximation, that is they assume that the target residues are in the same environments as the template residues if they belong to the same structural fold. In reality, even two closely homologous structures can have slightly different residue environments, especially in loop regions. This is one reason why Bryant's group use only conserved regions in threading (Bryant and Lawrence, 1993; Madej *et al*., 1995).

Ab Initio Prediction

Despite the great effort previously spent on comparative modelling and threading, there remains a large proportion of protein sequences with neither homologues nor clear folds detected. From the early 1970s, people began starting to look for ambitious *ab initio* algorithms that could directly attack the protein folding problem, that is to use supercomputers to explore the huge conformational space of protein molecules and find the pathways that lead proteins to their native conformations. The methods are based on the assumption that a protein molecule's native structure is the lowest free energy state among all its possible alternative conformations. This assumption has been demonstrated to be true by much experimental data, most famously the pioneering experiment of Christian Anfinsen. The attraction of the *ab initio* approach is that it not only promises to solve the protein structure prediction problem without being limited by the current protein structure database but it can also provide theoretical explanations of how proteins fold into their native structures – in other words the answer to the famous protein folding enigma.

From the 1970s, scientists from various fields, including biology, chemistry, physics, computer science and mathematics, have collaborated to develop all sorts of *ab initio* structure prediction methods and have published numerous papers. However, no significant progress was made over a very long period. In recent years, because of the rapid expansion of experimental data and the rapid increase in computer speeds, deeper insight has been gained to the protein folding problem and new algorithms have been developed that are beginning to show encouraging results in the blind protein structure prediction tests (Moult *et al*., 1999).

Figure 2 gives a schematic view of *ab initio* prediction algorithms (after Lin, 1996). The figure indicates that three components are essential for designing an *ab initio* algorithm, shown as the three dimensions in the figure. All the *ab initio* folding algorithms can be considered different combinations of the three components.

The first dimension in **Figure 2** is the protein model, which is used to characterize the protein molecules in the computer. This can be as complicated as the explicit atomic model in the classic molecular dynamics programs, in which all protein atoms and their related physical chemical properties (bond, order, length and angle, electronic charge, etc.) are explicitly described; or it can be a simple model like the simplified residues model, in which each residue is represented as a single particle in space. The lattice model represents the protein atoms or residues using discrete integer points in three-dimensional space, so the program is faster. Generally, the more complicated a model, the better it can describe the physical chemical properties of proteins, but also the slower the algorithm will be.

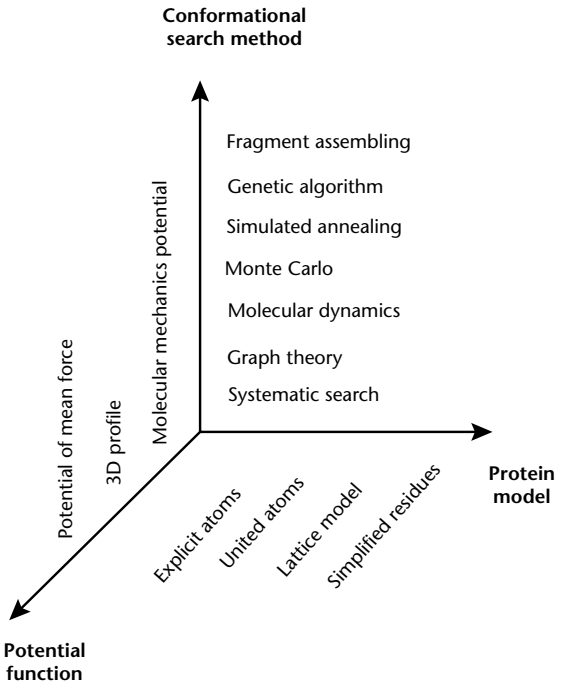

Figure 2 Schematic view of *ab initio* prediction methods (revised from Lin, 1996).

Potential function is the second dimension in **Figure 2**; this describes the physical chemical interactions both within protein molecules and between protein molecules and their environments. The ideal potential function is expected to rank the native conformation as the lowest free energy conformation among all possible alternatives. One of the most popular potential functions used in *ab initio* algorithms is the molecular mechanics potential widely adopted in molecular dynamics and molecular mechanics simulations, such as CHARMM (Brooks *et al.*, 1993), AMBER (Pearlman *et al.*, 1995) and GROMOS (van Gunsteren and Berendsen, 1990). Its general form is shown in eqn [2].

$$V(r_1, r_1, ..., r_N) = \sum_{\text{bonds}} \frac{1}{2} k_b (b - b_0)^2 + \sum_{\text{angles}} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \sum_{\substack{\text{improper} \\ \text{dihedrals}}} \frac{1}{2} k_\xi (\xi - \xi_0)^2 + \sum_{\text{dihedrals}} k\varphi [1 + \cos(n\varphi - \delta)] + \sum_{\text{pairs}(i,j)} \left[\frac{C_{12}(i,j)}{r_{ij}^{12}} - \frac{C_6(i,j)}{r_{ij}^6} + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \right] \quad [2]$$

The first term in eqn [2] is the bond stretch interaction along the covalent bond direction. It is represented by a harmonic function, in which b is the bond length and the values of the minimum energy bond length b_0 and force constant k_b are dependent on the specific bond type. The second term is the bond angle bending potential, which is a three-body interaction; θ, θ_0 and k_θ are the bond angle, minimum-energy bond angle and force constant. The four-body interactions fall into two categories: one is a harmonic potential to constrain the dihedral angle ξ, the other is a cosine potential that allows the dihedral angle φ to rotate 360°; k_ξ, k_φ, ξ_0, δ and n are the corresponding constants. The last summation term is the sum of two terms representing nonbonding interactions, which consist of the van der Waals potential and the electrostatic potential between atoms i and j. C_{12} and C_6 are the Lennard-Jones constants, r_{ij} is the distance between atoms i and j, and ε_0 and ε_r are the dielectric constant in vacuum and the relative dielectric constant in a medium.

The advantage of the molecular mechanics potentials is that they can explicitly characterize the physical chemical interactions in proteins at detailed atomic scale; but they are very slow to compute and also are not good for evaluating the solvent interactions, especially the important solvent entropy effect in protein folding. Thus, many of the latest *ab initio* folding algorithms prefer to use simple threading potentials as described earlier. The threading potentials, which are also called knowledge-based potentials, are derived from the current protein structure database and reflect either residue–residue distance distribution probabilities or residue-to-environment and residue-to-structure propensities.

The last dimension in **Figure 2** is the conformational search method, which is how the conformational space of proteins is explored to look for the lowest free energy conformation. Since proteins are long-chain biopolymers, they have a large number of internal degrees of freedom originating from both main-chain and side-chain dihedral angles. The simplest conformational search method is the systematic search. This divides each dihedral angle into a few discrete states approximately representing the local energy minima of that angle. One can then generate approximately all the possible conformations of the whole molecule by combining all the states of each dihedral angle. Because of the exponential increase in the number of combinations as the molecular size increases, it is actually impossible to use this method in any real protein systems.

The problem of exploring the conformational space of proteins is a typical combinatorial problem in computer science, which has been demonstrated to be NP-complete in complexity (Ngo and Marks, 1992). This means that no efficient algorithm is guaranteed to find the answer to the problem in a time bounded by a polynomial function of the protein size.

Present *ab initio* prediction algorithms use virtually every kind of advanced algorithm that has been used in

solving combinatorial problems, such as molecular dynamics (Duan and Kollman, 1998), Monte Carlo (Simons *et al*., 1999; Ortiz *et al*., 1999), genetic algorithms (Pederson and Moult, 1997), simulated annealing and graph theory methods. The molecular dynamics algorithm simulates the movement of the atoms of proteins and solvents based on classical Newtonian laws, and thus has a strong physics background. However, most of the latest *ab initio* prediction algorithms tend to use Monte Carlo algorithms or genetic algorithms because the most effective potential functions nowadays for *ab initio* prediction are knowledge-based threading functions, which in most cases are discrete and unable to calculate molecular forces for molecular dynamics simulations. Some workers have also tried to combine the molecular dynamics method with the Monte Carlo method in one algorithm, as well as combining different potentials (Zhang, 1999).

Fragment-assembling algorithms increase the conformational search efficiency by enumerating the limited number of possible structures for any given protein fragment. The possible candidate structures are selected on the basis of statistical analysis of the current protein structure database. Using these algorithms, it is not necessary to spend a great deal of time exploring the conformational space of every fragment; instead, whole protein conformations can be obtained by assembling the limited number of fragment conformations. As a result, the program can be fast enough to search the conformational space of small to medium-sized proteins currently using Monte Carlo or genetic algorithms. In addition to the speed advantage, the fragment-assembling algorithms can guarantee to give reasonable local structures, at least for the fragment structures selected.

In the history of protein structure prediction, the authors of *ab initio* algorithms have tended to overestimate the performance of their algorithms because of the lack of objective assessment methods. Starting from 1994, John Moult and his co-workers organized a series of conferences named CASP (Critical Assessment of techniques for protein Structure Prediction). The procedure of CASP is to first collect a number of protein targets whose structures are soon to be solved by X-ray crystallography, those targets are posted on the Internet, inviting predictors around the world to submit their predictions before the experimental structures become public. After the experimental structures are solved, the committee of CASP uses objective criteria such as the root mean square deviation between the predicted structure and the real structure to evaluate the success of all predictions.

The CASP3 results showed that several *ab initio* prediction groups have produced reasonably accurate models of protein fragments of up to 60 residues or so (Orengo *et al*., 1999; Simons *et al*., 1999; Ortiz *et al*., 1999), especially the fragment assembling algorithm (Simons *et al*., 1999).

Discussion

From the review in the previous sections, it can be seen that the comparative modelling method has become a very mature approach for protein structure prediction, while more recent advances in threading methods effectively extend the structure prediction scale to remote homologues. Finally, cutting-edge developments in software and hardware have brought *ab initio* algorithms very close to real application.

One of the latest developments related to protein structure prediction is the emergence of the structural genomics project in the post-human genomics era. After Celera Genomics and the public effort headed by NIH dramatically finished the human genome project (HGP) ahead of the expected timetable, scientists around the world started to collaborate on the structural genomics project (Sanchez *et al*., 2000). The idea is to classify all the proteins in the genome into homologous families and then to pick a representative sequence for each family to make experimental structures. Subsequently, the structures of all the sequences in the genome can be modelled using plain comparative modelling methods. In other words, all future protein structure prediction work would be comparative modelling. On the other hand, other structure prediction methods still can be useful in the future; for example, *ab initio* algorithms can still be used to study the theoretical basis of the protein folding problem.

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

Altschul SF, Madden TL, Schaffer AA *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

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

Bryant SH and Lawrence CE (1993) An empirical energy function for threading protein-sequence through the folding motif. *Proteins* **16**: 92–112.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S and Karplus M (1993) CHARMM: a program for macromolecular energy minimization, and dynamics calculations. *Journal of Computational Chemistry* **4**: 187–217.

Dayhoff MO, Schwartz RM and Orcutt BC (1978) A model of evolutionary change in protein matrices for detecting distant relationships. In: Dayhoff MO (ed.) *Atlas of Protein Sequence and Structure*, vol. 5, supplement 3, pp. 345–352. Washington, DC: National Biomedical Research Foundation.

Di Francesco V, Garnier J and Munson PJ (1997a) Protein topology recognition from secondary structure sequences: application of the hidden Markov models to the alpha class proteins. *Journal of Molecular Biology* **267**: 446–463.

Di Francesco V, Geetha V, Garnier J and Munson PJ (1997b) Fold recognition using predicted secondary structure sequences and hidden Markov models of protein folds. *Proteins* (supplement 1): 123–128.

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

Dunbrack RL Jr (1999) Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL. *Proteins* (supplement 3): 81–87.

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

Godzik A and Skolnick J (1992) Sequence–structure matching in globular proteins: application to supersecondary and tertiary structure determination. *Proceedings of the National Academy of Sciences of the USA* **89**: 12098–12102.

Henikoff S and Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the USA* **89**: 10915–10919.

Higgins DG and Sharp PM (1989) CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. *Gene* **73**: 237–244.

Jones DT, Taylor WR and Thornton JM (1992) A new approach to protein fold recognition. *Nature* **358**: 86–89.

Leach AR (1996) *Molecular Modelling: Principles and Applications*. Essex: Addison Wesley Longman.

Lin D (1996) *Knowledge-based Protein Fold and Folding Study*. PhD thesis, Peking University, p. 76.

Luthy R, Bowie JU and Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. *Nature* **356**: 83–85.

Madej T, Gibrat JF, Bryant SH (1995) Threading a database of protein cores. *Proteins* **23**: 356–369.

Moult J, Hubbard T, Fidelis K and Pedersen JT (1999) Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins* (supplement 3): 2–6.

Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology* **48**: 443–453.

Ngo JT and Marks J (1992) Computational complexity of a problem in molecular structure prediction. *Protein Engineering* **5**: 313–321.

Orengo CA, Bray JE, Hubbard T, LoConte L and Sillitoe I (1999) Analysis and assessment of *ab initio* three-dimensional prediction, secondary structure, and contacts prediction. *Proteins* (supplement 3): 149–170.

Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B and Skolnik J (1999) *Ab initio* folding of proteins using restraints derived from evolutionary information. *Proteins* (supplement 3): 177–185.

Pearlman DA, Case DA, Caldwell JW *et al*. (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computational Physics Communications* **91**: 1–41.

Pearson WR and Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the USA* **85**: 2444–2448.

Pearson WR (1990) Rapid and sensitive sequence comparison with PASTP and FASTA. *Methods in Enzymology* **183**: 63–98.

Pederson JT and Moult J (1997) *Ab initio* protein folding simulations with genetic algorithms: simulations on the complete sequence of small proteins. *Proteins* (supplement 1): 179–184.

Rost B and Sander C (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19**: 55–72.

Sanchez R, Pieper U, Melo F *et al*. (2000) Protein structure modeling for structural genomics. *Nature Structural Biology* **7** (supplement): 986–990.

Simons KT, Bonneau R, Ruczinski I and Baker D (1999) *Ab initio* protein structure prediction of CASP III targets using Rosetta. *Proteins* (supplement 3): 171–176.

Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology* **213**: 859–883.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology* **147**: 195–197.

van Gunsteren WF and Berendsen HJC (1990) Computer simulation of molecular dynamics: methodology, applications and perspectives in chemistry. *Angewandte Chemie. International Edition in English* **29**: 992–1023.

Wang ZX (1998) A re-estimation for the total numbers of protein folds and superfamilies. *Protein Engineering* **11**: 621–626.

Zhang H (1999) A new hybrid Monte Carlo algorithm for protein potential function test and structure refinement. *Proteins* **34**: 464–471.

Zhang H, Lai L, Wang L, Han Y and Tang Y (1997) A fast and efficient program for modeling protein loops. *Biopolymers* **41**: 61–72.

Further Reading

Leach AR (1996) *Molecular Modelling: Principles and Applications*. Essex: Addison Wesley Longman.

Eisenhaber F, Persson B and Argos P (1995) Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. *Critical Reviews in Biochemistry and Molecular Biology* **30**: 1–94.