Recent developments in structural proteomics for protein structure determination

Hsuan-Liang Liu¹ and Jyh-Ping Hsu²

¹ Department of Chemical Engineering and Graduate Institute of Biotechnology, National Taipei University of Technology, Taipei, Taiwan
² Department of Chemical Engineering, National Taiwan University, Taipei, Taiwan

The major challenges in structural proteomics include identifying all the proteins on the genome-wide scale, determining their structure-function relationships, and outlining the precise three-dimensional structures of the proteins. Protein structures are typically determined by experimental approaches such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. However, the knowledge of three-dimensional space by these techniques is still limited. Thus, computational methods such as comparative and de novo approaches and molecular dynamic simulations are intensively used as alternative tools to predict the three-dimensional structures and dynamic behavior of proteins. This review summarizes recent developments in structural proteomics for protein structure determination; including instrumental methods such as X-ray crystallography and NMR spectroscopy, and computational methods such as comparative and de novo structure prediction and molecular dynamics simulations.

Keywords:
Molecular dynamics simulations / Nuclear magnetic resonance spectroscopy / Review / Structural proteomics / Structure-function relationship / X-ray crystallography

Contents
1 Introduction .................................. 2056
2 Instrumental methods for structure determinations ....................... 2057
  2.1 X-ray crystallography ...................... 2057
  2.2 NMR spectroscopy ....................... 2059
3 Computational methods for structure prediction ........................ 2060
  3.1 Comparative and de novo structure prediction .......................... 2060
  3.2 Molecular dynamics solutions ........... 2063
4 Conclusions ................................ 2066
5 References ................................ 2066

1 Introduction

The term “proteome” was coined in 1994 by Marc R. Wilkins, vice president and head of bioinformatics at Proteome Systems in Sydney, Australia, to mean the protein complement encoded by a genome. Since then, the activities in proteomics can be broken down into three main categories: (1) identifying all the proteins made in a given cell, tissue, or organism; (2) determining how these proteins join forces to form networks akin to electrical circuits; and (3) outlining the precise three-dimensional (3-D) structures of the proteins in an effort to find their Achilles’ heels, i.e., where drugs might turn their activity on or off. Structural proteomics, the determination and prediction of atomic resolution 3-D structures of proteins on a genome-wide scale for better understanding their structure-function relationships, has now provided a new rationale for structural biology and has become a major initiative in biotechnology [1]. It usually involves research in biochemistry, bioinformatics, molecular biology, instrumental methods such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, and computational approaches. The activities involved in protein structure determination can be summarized in Fig. 1. The large num-
ber of protein structures will yield valuable information to the rules for predicting protein folding/unfolding mechanisms and understanding their biological functions.

Although recent advances in the fields of X-ray crystallography [2, 3] and multidimensional NMR spectroscopy [4] have allowed structural biologists to annotate the structures and biological functions of proteins by determining their atomic coordinates, they still suffer from some experimental limitations, such as the efficient and rational production of proteins with structural properties including high-throughput cloning and expression from multiple vectors in multiple host organisms, core domain identification using proteolysis methods, and the use of expression and detection tags. Therefore, computational approaches such as conformational search methods and molecular dynamics (MD) simulations may become alternatives capable of predicting the 3-D structures and understanding the folding/unfolding mechanisms of proteins. One of the advantages to use computational approaches in structural proteomics is to suggest the structures and biological functions of uncharacterized proteins solely based on structural homology to another protein with known structure and biological functions. Such predicted structural models may further provide valuable information for the better understanding of the putative activities or functions in protein–protein interactions and metabolic pathways in a given cell, tissue, and organism.

This review focuses on recent developments in structural proteomics for protein structure determination. Section 2 gives a brief background on instrumental methods such as X-ray crystallography and NMR spectroscopy; Section 3 provides details in computational methods including comparative and de novo structure prediction and MD simulations; and Section 4 outlines the concluding remarks of protein structure determination in structural proteomics.

2 Instrumental methods for structure determinations

Improvements in solving protein structures have been made by numerous innovations over the past decade, including solving the X-ray diffraction phase problem by multiple-wavelength anomalous diffraction (MAD) phasing [5]. These efforts together have resulted in dramatic accumulation of 3-D protein structures in the past few years (Fig. 2) due to the availability of genome sequences and technical advances in cloning and expression [6, 7], despite the fact that many of these structures are redundant (i.e., single amino acid mutations, ligand complexes, and duplications).

2.1 X-ray crystallography

To date, X-ray crystallography is still the most powerful technique for structure determination and analysis of proteins because it is capable of providing atomic coordinates of the whole assembly [8, 9]. Ever since the determination of the structure of myoglobin, these structural data have contributed tremendously to our understanding of structural biology at the atomic level [10]. Almost all proteins subjected to structural studies are expressed in heterologous systems. Recent developments in molecular biology have allowed high-throughput cloning and expression of proteins. However, overexpression in host cells may result in inappropriate PTM and incorrect folding, which usually forms inclusion bodies and insoluble aggregates, and thus must be discarded. Such limitations have been conquered by a number of strategies, such as using genes from different species, altering constructs, screening for solubility, and utilizing different cellular or cell-free expression systems. Furthermore, several novel refolding techniques such as refolding
chromatography based on molecular chaperones that are part of the cellular machinery responsible for folding nascent protein [11] and the designed small-molecule agents to assist the refolding process [12], have been developed to recover these proteins.

Once the proteins are correctly expressed and purified, the next step is to form crystals of sufficient quality to collect high-resolution data for structure determination. Crystallization is usually regarded as a slow, resource-intensive step with low success rates. However, much of the failure in this step can be attributed to poor protein quality. Therefore, the use of biophysical methods such as dynamic light scattering to assess the quality of a protein is a key step before performing crystallization experiments [13]. In addition, it is usually necessary to screen a wide range of conditions such as pH, temperature, salt and protein concentrations, and cofactors, in this step. Particularly for proteins with low yields, the ability to screen more conditions at the required protein concentration becomes critical. Crystallization has benefited enormously from automation and technologies allowing the use of small protein sample volumes [2] over the past few years. Meanwhile, crystallization robots have contributed to increasing the efficiency of screening crystallization conditions. Recently, the so-called second-generation crystallization robots have been developed to allow many thousands of experiments to be carried out in parallel. These robots significantly reduce the need for large amounts of protein samples by using nanoliter-sized drops to screen for the optimal conditions. In addition, they also provide image capture and analysis systems in place, which monitor the drops automatically and look for the presence of crystals using edge-detection methods [2].

As mentioned above, high-throughput crystallography has been facilitated by improved phasing and model building methods, decreased sample requirements through miniaturization, as well as robotization and automation from the crystallization stage to structure determination [14]. Thus, the process from genes to crystals to drug candidates is now possible and has been intensively executed in academic research and commercial groups [15]. Beyond the high-throughput capabilities for protein expression, crystallization, image analysis for automatic crystal detection, and structure determination, a late stage rate-determining step, is the manual intervention required for crystal mounting and alignment in the X-ray beam. A novel method for automatic mounting, optical crystal alignment, and data collection has been reported recently [16]. It exhibits the same range of accuracy as the manual process and allows data collection using conventional X-ray systems. In addition, the number of high-brilliance synchrotron X-ray beam lines dedicated to macromolecular crystallography has increased significantly, and thus data collection times at these facilities can be dramatically reduced.

Figure 2. The growth of total available protein structures deposited in PDB (figure downloaded from PDB website: http://www.rcsb.org/pdb).
The technological advances in X-ray crystallography, over the past decade, have resulted in rapid growth of protein structure data. However, crystal structures of multi-component systems and membrane proteins are still limited, providing key challenges for these new approaches. Nevertheless, it is highly anticipated that these novel approaches will continue to generate protein crystal structures at unprecedented rates and drive a new wave of structure-based drug discovery [15].

2.2 NMR spectroscopy

Although X-ray crystallography is regarded as the potential workhorse for structural proteomics, the throughput of structure determination using it remains unclear. In contrast, NMR spectroscopy does not require crystals and thus samples appropriate for structure determination can be identified in a relatively short period of time. Moreover, NMR experiments can be carried out in aqueous solution under conditions similar to the physiological conditions in which the protein normally functions. These particular features have provided insights into structure–function relationships for a large number of proteins. However, structure determination by NMR is currently limited by size constraints and lengthy data collection and analysis times. Nevertheless, NMR spectroscopy still plays a significant role in structural proteomics even with its current limitations [17].

The procedure for NMR solution structure determination can be dissected into six major parts. (1) A suitable sample, usually ~500 µL of a 1 mM protein solution is carefully prepared. If the molecular mass of a protein exceeds ~10 kDa, enrichment with 13C and 15N isotopes is required to resolve spectral overlap in 1H NMR spectroscopy. (2) This sample is subsequently used to record a set of multi-dimensional NMR experiments, which provide the NMR spectra after suitable data processing. (3) Determine the complete sequential NMR assignments such as the measurement of resonance frequencies of the NMR-active spins in the protein. (4) The resulting conformation-dependent dispersion of the chemical shifts is a prerequisite for deriving experimental constraints from various NMR experiments for the NMR structure calculation. (5) Iterations involving structure calculations and identification of new constraints are carried out until the overwhelming majority of experimentally derived constraints are in good agreement with a bundle of protein conformations representing the NMR solution structure. (6) Finally, the NMR structure can be refined using conformational energy force fields [4].

As conformational variations in the bundle of structures reflect the precision of the NMR structure determination, the accuracy of protein structures determined by NMR is strongly dependent on the extent and quality of data obtained. In addition, the sensitivity of the acquired NMR data relies strongly on the performance of the NMR probe, a sophisticated electronic device used to detect NMR signals. Although atomic coordinates in high-resolution crystal structures are more precisely determined than in the corresponding NMR structures, the crystallization process may select a certain subset of conformers present under particular conditions. Recently, various approaches towards the automated structure determination from NMR spectra, such as the NOE-Jigsaw [18] in which sparse and unassigned NMR data can be used to reasonably and accurately assess secondary structure and align it, have been developed. The information obtained is thus useful in quick structural analysis to assess folds before full structure determination and analysis as a complement to fold prediction approaches [14].

In addition, NMR is particularly valuable in structural genomics efforts for rapidly characterizing the foldedness of specific protein or RNA constructs [4]. In principle, there are correlations between such foldedness criteria and crystallizability, so that data from a high-throughput NMR screen might directly support efforts to generate samples for crystallographic analysis [4].

As mentioned, the major challenge for NMR spectroscopy is the reduction of the data collection time required for structure determination. Besides, the development of NMR experiments that allow matching of instrument time investments to the minimum time required for measuring the chemical shift data is another challenge. They can be overcome by reduced dimensionality experiments [19, 20], with simultaneous frequency labeling of more than one atom type in indirect dimensions. Another advance involves partial deuteration [21], which provides samples that can be studied with improved S/Ns that result from their sharper linewidths and longer transverse relaxation times. These technologies together provide the basis for high-throughput NMR, and are particularly valuable for samples with limited stabilities and low solubilities. A novel spectroscopic concept, transverse relaxation optimized spectroscopy (TROSY) based on selection of slowly relaxing NMR transitions, also can provide significant sensitivity enhancement for large proteins [22, 23].

Another important issue of development involves automated analysis of NMR data. The design of cryogenic probes will certainly assist in reducing the bottleneck of time required for data collection in NMR-based structural proteomics. In addition, some automated resonance- and nuclear Overhauser effects (NOE)-assignment programs, such as NOAH [24], AUTOASSIGN [25], ARIA [26], ANSIG [27], TATAPRO [28], and SANE [29], hold great promise in reducing the amount of time in structure determination. These recent developments provide automated analysis of NMR assignments and 3-D structures of proteins up to 200 amino acids [30]. However, general methods for automated analysis of side chain resonance assignment are not yet well developed, and there are as yet no examples of completely automated protein structure determinations. Nevertheless, NMR spectroscopy can provide structural and biophysical information that is complementary to X-ray crystallography, and these two methods will continue to play synergistic roles in structural proteomics.
3 Computational methods for structure prediction

3.1 Comparative and de novo structure prediction

Computational methods usually involve comparative structure prediction and de novo structure prediction. Comparative methods usually follow four basic steps: (1) finding a template; (2) sequence-template alignment; (3) model building; and (4) model assessment. Successful predictions based on comparative models have been reviewed by Baker and Sali [31]. It usually entails the use of information from a variety of sources including pattern analysis, X-ray diffraction data, and physical forces/energy states. The combined use of these sources of information would predict the probable atomic coordinates of proteins in space. If there is no empirically determined structure with at least 30% sequence similarity to the target sequence, then there may be no template available that is suitable for reliable comparative modeling. Thus, the potential accuracy of a comparative model is positively related to the percentage of sequence identity on which it is based. High-accuracy comparative models are based on more than 50% sequence identity to their templates. They tend to have about 1 root-mean-square (RMS) deviation for the main-chain atoms, which is comparative to the accuracy of a medium-resolution NMR structure or a low-resolution X-ray structure [31]. The errors of the homology models occur primarily in the area of mistakes in side-chain packing, small shifts or distortions in the core main-chain regions, and inserted loops with no matching sequence in the solved structure. Furthermore, sequence alignment mistakes may occasionally take place [32]. Although there is a wide range of applications of protein structure models, the scope of modeling is still limited to the content of the protein structure database. According to Vitkup et al. [33], providing templates for 90% of all protein domain families, including membrane proteins, will require the empirical solution of about 16,000 new domain structures. This is well in excess of the structural information presently in the protein data bank (PDB), but may be achievable within a decade as a result of advancements in high-throughput crystallography and NMR spectroscopy. The homology models may be further investigated in a variety of ways, such as site-directed mutagenesis, ligand docking experiments, and protein–protein interactions.

An example of using comparative methods to construct the homology model of the pore-loop domain of the voltage-gated potassium channel Kv1.1 from human Homo sapiens based on the crystal structure of the bacterial potassium channel from Streptomyces lividans (KcsA) [34] is shown in Fig. 3 [35]. The built model shares a similar fold as in KcsA, particularly in the selectivity filter, a signature sequence forming the sequential K⁺ ion binding sites. In contrast, most of the structural variations occur in the turret as well as in the inner and outer helices. The homology models of these Kv channels provide particularly attractive targets for further structure-based studies. Molecular docking experiments of a newly solved scorpion toxin, Tc1, from Tityus cambridgei [36] were subsequently performed towards these channels to validate the built structure and the results are illustrated in Fig. 4 [37]. The side chain of Lys14 was found to be directly involved in the electrostatic interactions with the first K⁺ ion-binding site in the selectivity filter. In addition, Fig. 5 shows that Tc1 exhibits stronger binding affinity towards Kv1.1 than KcsA due to stronger electrostatic and hydrophobic interactions between Tc1 molecule and the outer vestibule of Kv1.1 [38].
Figure 4. Ribbon representations of the top and side views of the pore loop domain of (A) KcsA and (B) Kv1.1 after docked by Tc1 molecule [37], which is illustrated in CPK. The four subunits of each channel, indicated as I, II, III, and IV, are shown in blue, orange, green, and red, respectively. The locations of the outer helix, turret, pore helix, selectivity filter (represented as sticks with the backbone carbonyl oxygen atoms in red), and inner helix [84] are indicated. One of the four subunits of the two channels is omitted in the side views for clarity. The five Lys residues of Tc1 are labeled and colored in yellow. The rest of the residues in Tc1 are colored in purple. Figure reproduced with permission.

Unlike comparative methods that are limited to protein families with at least one known structure, de novo prediction is free of such a restriction. Starting on the premise that the native state of a protein is at the global-free energy minimum, the search of conformational spaces for tertiary structures that are particularly low in free energy for a given amino acid sequence is possible. Recent advance in this area introduces a toolkit, ROSETTA, for ab initio protein structure prediction [39]. ROSETTA is based on the assumption that the distribution of conformations sampled for a given nine residue segment of the chain is reasonably well approximated by the distribution of structures adopted by the sequence in known protein structures [40]. Fragment libraries for all possible three- and nine-residue segments of the chain are extracted from the protein structure database by using a sequence profile comparison method. The conformational space defined by these fragments is searched by using a Monte Carlo procedure with an energy function that favors compact structures with paired β-strands and buried hydrophobic residues. The output structures are by con-
struction consistent with the local conformational biases inherent in the sequence and have low free-energy non-local interactions by virtue of the Monte Carlo optimization procedure. For each query sequence a large number of independent simulations are carried out by starting from different random number seeds. The resulting structures are clustered, and the centers of the largest clusters are selected as the highest confident models. Thus, it provides modeling support for different design domains, employing semantics and syntax appropriate for each. Structural genomics of single proteins or their domains, combined with protein structure prediction, may help scientists to efficiently characterize the structures of large macromolecular assemblies. Results from CASP (III, IV, and V) protein structure prediction experiments demonstrated that considerable progress has been made towards predicting protein structure from its primary sequence information alone [41–43]. The primary improvements to the method since CASP III fall into three classes. The first class consists of improvements in the basic simulation method. The second class of improvements consists of filters that eliminate nonprotein-like conformations from large sets of simulated structures and increase the frequency of native-like conformations. The third class of improvements involves the simultaneous clustering of conformations generated independently for several sequences related to a given target sequence.

Comparative modeling is based on the observation that proteins with similar sequences almost always share similar structures. Currently, about 30% of known sequences have sufficient sequence similarity to a known structure for current comparative modeling methods [44]. About one-third of these sequences are similar over <80% of their length; consequently, complete 3-D models cannot be generated by comparative methods alone [33]. Recently, a novel method based on the successful de novo structure prediction method ROSETTA [45, 46] for modeling structurally variable regions has been proposed [44]. Traditionally, loop modeling is defined as the problem of constructing 3-D atomic models for short protein segments corresponding to loops on the protein surface that connect regular secondary structure elements. Loop modeling methods primarily differ in the method of conformation generation and in the evaluation or scoring of alternative conformations. Algorithms can be generally grouped into knowledge-based methods, de novo or ab initio strategies, and combined approaches. The knowledge-based approach uses the database of experimental protein structures as a source of loop conformations. In the de novo approach, loop conformations

Figure 5.
are generated by a variety of methods including MD [47, 48], simulated annealing [49, 50], exhaustive enumeration or heuristic sampling of a discrete set of ($\phi$, $\psi$) angles [51, 52], random tweak [53, 54], or analytical methods [55, 56]. The ROSETTA-based method described by Rohl et al. [44] is a novel approach for combining database-derived conformations and de novo prediction for loop modeling. The fragment assembly strategy used by ROSETTA is currently perhaps the most successful method for de novo structure prediction, and it may be particularly well suited for modeling structurally variable regions in proteins.

Methods for protein structure prediction offer some hope for narrowing the gap between the number of known protein sequences and the number of experimentally determined protein structures. Fold recognition methods can be very effective despite their slowness, the requirement for human intervention to interpret the results, and the inaccuracy of sequence-structure alignments produced. Threading methods were originally developed to recognize pairs of proteins, which have no obvious similarities in sequence yet have similar folds. Recently, it has become clear that in fold recognition it is useful to distinguish between pairs of proteins, which are homologous and those, which are analogous. Where no evolutionary relationship is believed to exist between two structurally similar proteins, threading would be the only applicable method for identifying this type of relationship. Usually, the method for threading can be divided into three stages: (1) alignment of sequences; (2) calculation of pair potential and salvation terms; and (3) evaluation of the alignment using a neural network. One of the possible applications for a fold recognition method might be to prioritize these gene products for X-ray crystallographic experiments. Thus, it can be hoped that a very large fraction of proteomes might be amenable to large-scale structure prediction in the near future.

3.2 Molecular dynamics simulations

Though their folding time may not be a subject of active refinement of evolution, proteins need to adapt well-defined structures soon after being synthesized and transported to

Figure 5. Molecular surfaces of the complexes of (A) KcsA/Tc1 and (B) Kv1.1/Tc1 with Lys14 of Tc1 facing towards the selectivity filters of these two channels [38]. Top views of the P-loop regions of KcsA and Kv1.1 after removing Tc1 molecules are shown on the left of (A) and (B), respectively. The dashed-yellow lines indicate the binding locations of Tc1 molecules on the channel surfaces after docking. The four Asp residues forming the hydrophobic interactions with Tc1 from each subunit are also labeled. Top and side views of the Tc1 molecules after docking towards KcsA and Kv1.1 are shown on the top-right and bottom-right of (A) and (B), respectively. The five Lys residues are labeled. Figure reproduced with permission.
their designated locations to perform their biological functions. Under the right physiological conditions, proteins can fold into and subsequently maintain the well-defined structures through the delicate balance of enthalpy and entropy [57], weak interactions including van der Waals, electrostatic, and hydrogen bonding forces, and a balance between protein intramolecular interactions and the interactions with solvent. Therefore, in addition to the task of protein structure determination in structural proteomics, understanding the mechanism of protein folding/unfolding is often referred to as the second half of genetics. There is a growing body of evidence indicating that while folding can proceed along different routes, some paths are more populated than the others [58–61]. Theoretical approaches to predict the stable folded structure or the process of protein folding/unfolding fall into three categories: (1) statistical approaches; (2) conformational search methods; and (3) dynamics simulations [62].

Statistical approaches relate amino acid sequence to the known 3-D structures and are reasonably successful at predicting secondary structure elements [63]. Conformational search methods can, in principle, yield the relative population of different conformers. A range of conformations is generated and an energy function is used to discriminate between them. Dynamics simulations have been intensively adopted to characterize particular folded states in solution [64, 65], but it is generally assumed that simulations involving physical force fields, atomic degrees of freedom, and explicit solvent cannot be used to predict protein folding in solution. The tendency has been, instead, to turn to simplified models or representations [66, 67], which have been widely applied to understand the physical principles governing the folding/unfolding processes and will continue to play important roles in the endeavor. Unfortunately, none of these simplified approaches has met with clear success. The challenge is to reproduce the free-energy surface of the molecule in solution with sufficient fidelity. Although simplified models allow greater sampling of conformational space than implicit model, it may sample many regions with inappropriate weights.

Computational studies of protein structures and folding/unfolding processes have come of age. Among the early successes were the lattice models, which allow efficient sampling of computational space. When designed properly, these models can give a well-defined global energy minimum and can be applied to structure prediction with encouraging results [68]. There are two types of lattice models aimed at two distinct objectives. The first category was designed to understand the basic physics governing the protein folding/unfolding mechanisms [69]. The second one, originated from the work by Miyazawa and Jernigan [70], was geared towards realistic folding of real proteins as templates by statistical sampling of the available structures [71] and is often referred to as statistical or knowledge-based potentials. In the past, the most widely used approaches in protein structure prediction have been based on residue-level methods with typically statistical potentials obtained from the structural database (i.e., PDB). A growing trend in the community has been the development of atomic-level statistical potentials [72–74] in attempts to improve the accuracy. It has been suggested that solvent plays a role as the lubricant prior to reaching the native state [75] and ejection of a solvent molecule from the interior may contribute a non-trivial portion to the free-energy barriers [76]. Therefore, the accuracy of these models can be further improved with the inclusion of the solvent effect [77, 78].

The parameters used in the all-atom representation of both solvent and protein, are obtained through high-level quantum mechanical calculations on short peptide fragments. Such an approach assures the generality and allows further refinement upon the availability of more accurate quantum mechanical methods and upon the need for such an improvement. These models have been widely implemented to study the unfolding processes of small proteins by temperature jump technique [75, 79–82], changing solvent condition [83, 84], applying external forces [85, 86], or pressure [83]. They can be also applied to study the ligand recognition [87–89] or the stabilization effect of secondary structures [90]. Figure 6 shows the snapshots from a series of 200 ps MD simulations at various temperatures [82] to investigate the unfolding mechanism of the catalytic domain of Aspergillus awamori glucoamylase. The unfolding of this domain was suggested to follow a putative hierarchical manner, in which the heavily O-glycosylated belt region from residues T440 to A471 acted as the initiation site, followed by the α-helix secondary structure destruction, and then the collapse of the catalytic center pocket. In addition, MD simulations have also been applied to determine the role of alcohols with various aliphatic chain lengths on the secondary structure integrity of monomeric melittin [84]. The results of secondary structure analysis according to DSSP [91] (Fig. 7) indicate that the weaker dielectric constant of longer aliphatic chain length of alcohol possibly reduces the hydrogen bonding between amide protons and surrounding solvent molecules and simultaneously promotes the intramolecular hydrogen bonding in melittin. Besides unfolding simulations, limited refolding simulations were also attempted with partially unfolded structures generated from the unfolding simulations and considerable fluctuations were observed [92]. These refolding experiments have also identified the transition states in the vicinity of the native state [93].

Proteins can fold from fully unfolded states to the native state by going through many intermediate states of varying degrees of stability. Thus, studying the mechanism of protein folding is regarded as the study of these intermediate states, the relationship between them, and the relationship between them and the native state [94, 95]. However, a major difficulty in experimental studies of intermediate states is the transient nature of these states, necessitating fast techniques [96–98]. In contrast, although it is still unrealistic to expect the protein to reach the native state during the simulation time scale by current computer power, this time scale appears to be sufficient to observe some marginally stable
intermediates. Previously, Kazmirski and Daggett [99] have successfully sampled various stable intermediate structures of hen egg-white lysozyme during two sets of 9 ns MD unfolding simulations. Their results suggested that the kinetic intermediate may be made up of distinct, but rapidly interconverting, partially folded conformations distinguished primarily by differences in secondary structure packing. Furthermore, the free energy of the intermediate state has been proved to be the lowest among all the states sampled during the simulation [100]. These results strongly suggest that the kinetics, not the force-field artifacts, is the barrier for the simulation to reach the native state.

Another approach to simplify the protein-folding problem is the building block folding model, in which protein fragments are produced either experimentally [101, 102] or computationally [103–106]. Whether these fragments can form independent folding entities is further determined [107]. In general, there is a correspondence in the regions of the polypeptide chains being cleaved by these two methods, despite the difference in their premises [108]. Therefore, proteolytic enzymes may be used as reliable probes of protein structure, dynamics, and folding pathways. Furthermore, the consistency with the experimental cleavages appears to provide a validation of the building block folding model, where the independently folding hydrophobic units are obtained through hierarchical assembly of these conformationally fluctuating building blocks [109, 110]. As a result, this model is in line with the proposal that protein folding is a hierarchical event [111, 112], where parts constituting local minima of energy fold first, with their subsequent association and mutual stabilization to finally yield the global fold.

With the promises of more accurate models and considerably faster computer speed, together with the advancement of experimental approaches, such as mutagenesis [113], hydrogen exchange experiments [114, 115], atomic force microscope (AFM) [116], solution structure technique [117, 118], and ultrafast kinetic experiments [119, 120], the goals of predicting protein structures from their primary sequences and understanding of the protein folding/unfolding mechanisms should be achieved in the near future. Hopefully, the diversity of protein structures and the complexity of the in vivo folding process can be dealt with by a combination of experimental approaches and further simulation work.

Figure 6. Snapshots of the catalytic domain from Aspergillus awamori glucoamylase at 50, 100, 150, and 200 ps at (A) 300 K; (B) 400 K; (C) 600 K; and (D) 800 K during 200 ps MD simulations [82]. The N- and C-terminus are indicated in the upper starting structure. The upper and lower starting structures at t = 0 indicate the numbering and the corresponding colors of the 13 α-helices, respectively. The same coloring was applied to each snapshot. Figure reproduced with permission.
4 Conclusions

Structure prediction methods are likely to play a key role in the generation of full structural proteome maps. As the number of structural templates increases, a vast experimental pool of data will be available for future machine learning experiments and hopefully lead to a rapid improvement in the accuracy and speed of such efforts. Furthermore, establishing protein interaction networks is crucial for understanding cellular operations. 3-D structures can be used to interrogate the whole interaction network to validate and infer molecular details for the interactions proposed by other approaches. We can expect that the determination and prediction of the 3-D structures and the structure–function relationship of proteins for studying protein–protein interactions will become increasingly important in the post-genomic era.

The authors gratefully acknowledge the financial support from the National Science Council of Taiwan (NSC-93–2214-E-027-001).

5 References
