Molecular Cell, Vol. 20, 811–819, December 22, 2005, Copyright ©2005 by Elsevier Inc. DOI 10.1016/j.molcel.2005.12.005

Protein Structure Prediction: Inroads to Biology

Review

Donald Petrey[1] and Barry Honig[1,*]
[1] Howard Hughes Medical Institute
Department of Biochemistry and Molecular Biophysics
Center for Computational Biology and Bioinformatics
Columbia University
1130 St. Nicholas Avenue, Room 815
New York, New York 10032

In recent years, there has been significant progress in the ability to predict the three-dimensional structure of proteins from their amino acid sequence. Progress has been due to new methods to extract the growing amount of information in sequence and structure databases and improved computational descriptions of protein energetics. This review summarizes recent advances in these areas and describes a number of novel biological applications made possible by structure prediction. Despite remaining challenges, protein structure prediction is becoming an extremely useful tool in understanding phenomena in modern molecular and cell biology.

Introduction

The prediction of the three-dimensional structure of a protein from its amino acid sequence is a challenge that has fascinated researchers in different disciplines for many years. The potential impact of significant advances in structure prediction is enormous, since we already have ample evidence of the importance of three-dimensional structure information in so many areas of biology. As an example, a recent paper used predicted structures of retroviral matrix domains, together with an analysis of known structures, to establish the mode of interaction of the entire matrix domain family with membrane surfaces (Murray et al., 2005). More generally, models of members of protein families that are derived from experimentally determined structures make it possible to deduce family-based structure-function relationships that are not evident from a small number of currently available representative structures of family members. The availability of such models can also reveal specificity differences within families, thus significantly expanding the range of questions that can be addressed based on the original experimentally determined structure. Models can also be used as a basis for identifying the function of individual proteins, much in the way this is accomplished with experimentally determined structures. However, despite the enormous potential impact of protein structure prediction, the extent to which models that are being generated today can be used with confidence in different applications is unclear. This review will address this question in the context of a discussion of the modeling process, with particular emphasis on a description of current research activity, and will also include a number of examples of specific biological applications of structure prediction.

Structure prediction is often divided into three areas: ab initio prediction, fold recognition, and homology modeling. The distinction between these areas is usually based on the extent to which information in sequence and structural databases is used in the construction of a model. Ab initio prediction in its purest form makes no use of information in databases (see e.g., Nanias et al. [2005]), and the goal is to predict the structure of a protein based entirely on the laws of physics and chemistry. The term ab initio or de novo is also applied to the prediction of the structure of proteins for which there is no similar structure in the Protein Database (PDB) but where local sequence and structural relationships involving short protein fragments, as well as secondary structure prediction, are incorporated into the prediction process (Bradley et al., 2005). Fold recognition, often referred to as "threading," corresponds to the case where one or more structures (templates) similar to a given target sequence exist in the PDB but are not easily identified. Here the main challenge is to find the best set of templates, but, in general, it will be difficult to build an accurate model. At the current stage of the technology, the most accurate models are obtained when a single template can be found in the PDB that has a high level of sequence similarity to the target protein. The process of building a model from such a template is referred to as homology or comparative modeling.

With the exception of pure physical chemical approaches, essentially all protein structure prediction methods rely on the identification of templates that may range in length from relatively short fragments, as in de novo methods, to entire proteins, as in homology modeling. Thus, they may all be thought of as involving "template-based protein structure prediction." In this review, we will use the term "protein structure prediction" to refer to all template-based methods. Our focus will be on methods that rely heavily on a small number of clearly defined templates since, at this point in time, these are the only ones that offer a high probability of being accurate enough to enable the identification of biological function.

The existence of templates that may cover only some regions of a protein raises new questions as to the nature of evolutionary relationships between proteins. Indeed, an important spin-off of structural biology has been the discovery of new relationships between amino acid sequences and protein structures, and among different protein structures. Methods to superimpose three-dimensional structures have been developed (Kolodny et al., 2005b), and it has been recognized that many proteins may have evolutionary relationships that are evident from structure but not from sequence. Widely used databases such as SCOP (Lo Conte et al., 2000) and CATH (Pearl et al., 2005) divide proteins into discrete families based on simple sequence relationships, superfamilies based on structural and function similarity and less-obvious sequence relationships, and fold based entirely on structural similarity. The assumption has been that proteins in different folds evolved independently, but is this generally the case?

*Correspondence: bh6@columbia.edu

Indeed, there is a great deal of ambiguity in the definition of a fold, and an argument can be made that fold space should be viewed as continuous (Shindyalov and Bourne, 2000; Yang and Honig, 2000). A related finding is that, given a target sequence, it is almost always possible either to find a protein that is similar to a target structure in overall topology (Kihara and Skolnick, 2003), or there is likely to be a set of proteins that have regions similar to all or part of the target (Szustakowski et al., 2005; Zhang and Skolnick, 2005). There is also evidence of sequence relationships between proteins that have different folds (Friedberg and Godzik, 2005). Thus, the problem of structure prediction appears dependent in part on the ability to identify sequence and structure relationships between proteins that, at least based on existing classification schemes, might not be expected to be related.

A significantly improved ability to detect such relationships may well be one of the important by-products of the various structural genomics initiatives that have been funded in different countries around the world. An important goal is to find a structural representative for as many protein sequence families as possible, so that homology models for other family members can be built, or at least so that an overall folding topology can be assigned to each member. That structure prediction is an essential component of structural genomics initiatives implies large-scale acceptance of the value of template-based structure prediction.

A number of reviews of structure prediction methods have recently been written (Ginalski et al., 2005; Marti-Renom et al., 2000; Sanchez and Sali, 1997). Also, the descriptions of the outcomes of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments are good sources of information about current structure prediction methods (Moult et al., 2003). CASP is a blind test of structure prediction methods in which the community of predictors spends a summer making predictions and then submits them for evaluation based on the experimentally determined structures. Many relevant tools for structure prediction are available online (see e.g., the yearly database issue of *Nucleic Acids Research* [Fox et al., 2005]). Our goal in this review is to provide an overview to the nonexpert of the structure prediction process, including a discussion of sources of error, bottlenecks, and recent advances that pertain to each stage. We also provide examples of various biological applications of structure prediction, particularly homology modeling, which we believe point to the nature of the applications that we may increasingly expect to see in the years to come.

The Structure Prediction Process

Template-based structure prediction can be thought of as involving six stages: (1) identification of related sequences with known structure (templates); (2) alignment of the target sequence to the template structure(s); (3) building an initial model for all or part of the target sequence based on the structure of the templates; (4) ab initio modeling of side chains and loops in the target that are different than the template; (5) model refinement, where the structure of the model is allowed to adopt a conformation different than that of the original model; and (6) model evaluation. Each of these stages is the subject of significant research activity, with the problems that are being addressed ranging in nature from the use of bioinformatics techniques to analyze sequence and structure to the physical-chemical study of the conformational energies and dynamics of proteins. In the following sections, we describe the tools currently used or being developed at each stage.

Template Selection

The first step in structure prediction invariably involves the use of some sequence alignment method to identify a statistically significant relationship between the target sequence and one or more possible templates. There are a variety of methods used for this purpose that have been reviewed elsewhere (Pearson and Sierk, 2005). The field has evolved in clear stages, differing primarily in the way in which the template sequences are represented. Currently, the most widely used template selection methods involve the representation of templates as profiles, as in PsiBlast (Altschul et al., 1997) or hidden Markov models (HMMs) (Eddy, 1998). In a profile or HMM, each position represents not a single amino acid, as in standard sequence alignments, but a collection of features that are obtained from the multiple sequence alignment of proteins that bear a clear evolutionary relationship. Profiles and HMMs, therefore, are a more accurate representation of the variability that can occur at individual positions of a protein sequence. This results in more sensitive detection of remote homologs (Marti-Renom et al., 2004).

The set of features included in a profile or HMM varies greatly. In their simplest form, each position in a sequence is simply the probability of each amino acid appearing at that position (Altschul et al., 1997). Structural features of the template and predicted structural features of the target, for example, the location of secondary structure elements, can also be included in a profile. In each case, one is matching known features of the template to predicted features (e.g., surface accessibility) of the target. Another way to incorporate structural information into what is essentially a linear description of a sequence is to use structure-based sequence alignments, in which residues are assumed to correspond if they occupy the same position in space in two different structures (Kelley et al., 2000; Tang et al., 2003). This procedure makes it possible to construct a profile for regions of a sequence even where there is no detectable sequence relationship, a situation in which it is difficult to construct a standard multiple sequence alignment. Fold recognition methods also use structural information about the template to obtain an alignment, generally through some measure of the fitness of each residue in the target sequence to be located in a particular environment in the template structure (Eisenberg et al., 1997; McGuffin and Jones, 2003; Zhou and Zhou, 2004). This measure might include, for example, the propensity of a residue for being buried/exposed or to be part of a particular secondary structure element. Fold recognition methods are frequently combined with profiles and HMMs.

The observed local similarities between ostensibly unrelated proteins (Bystroff and Baker, 1997) have led to the development of methods that use short fragments (<20 residues) to represent structural information in the

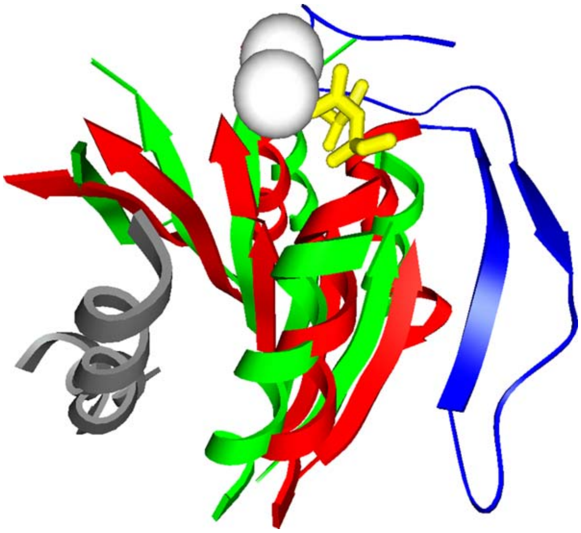

Figure 1. Using Remote Homology to Make Predictions

The figure shows a structural alignment of two proteins Spo0F (PDB code 1srr, chain A) and SurE (PDB code 1j9l, chain B). The two proteins share a large region of structural homology shown in green (SurE) and red (Spo0F). The differences between the two proteins include a long β hairpin insertion present in SurE but not Spo0F (blue), a structurally equivalent but circularly permuted helix (gray), and an 80 residue C-terminal extension (not shown for simplicity).

database. A model is built for a target sequence by finding all fragments that are similar to some part of the target sequence and combining these fragments to produce a compact fold (Simons et al., 1999). Such methods can produce accurate models where methods that attempt to find a single template with overall similarity to the complete target structure have failed (Bradley et al., 2005). Fragment assembly has also been combined with fold recognition. With these methods, fragments from models based on multiple templates that bear some overall relationship to the target are combined. The fragments that are selected to go into the final model are chosen based on their stability as measured by a simplified "scoring function" (Zhang and Skolnick, 2004).

The ultimate choice of a template depends on the statistical significance of the alignment score between two sequences, the proper evaluation of which is a subject of ongoing research (Pearson and Sierk, 2005). It should be recognized that there is an inherent shortcoming in all sequence alignment methods, since they implicitly assume that complex three-dimensional thermodynamic relationships can be mapped onto a linear sequence. It is to be expected, then, that measures of statistical significance will not always accurately reflect true structural or evolutionary relationships. As a result, it is possible for two proteins that share many structural similarities to have alignment scores that are in the range of what would be expected to be observed by chance. This can occur for many reasons, for example, when there are large structural differences between two proteins that complicate the identification of meaningful similarities in other regions.

Figure 1 contains an example where the statistical significance of the sequence relationship between two proteins, the calcium-dependent phosphatase SurE and sporulation response regulator (Spo0F), is marginal. The two proteins belong to different SCOP and CATH folds, but, nevertheless, SurE would serve as a good template for Spo0F. However, accurate modeling of Spo0F would require recognizing all the differences between the two proteins (see Figure 1). That the proteins might actually be evolutionarily related is suggested by the fact that both bind calcium (white sphere) with a structurally and sequentially homologous residue (yellow). Recognizing how to identify such apparently remote relationships in advance of knowing the two structures constitutes a significant research challenge.

Sequence-to-Structure Alignment

It is often the case that an alignment procedure is successful in identifying the correct template but that the target-template alignment is still not completely "correct" (Venclovas, 2003). In cases such as these, the correct alignment is "suboptimal" as a one-dimensional alignment in the sense that it has a lower alignment score than one or more incorrect alignments. (The correct alignment is taken to be the one that produces the best model, i.e., the alignment that most accurately identifies residues in the target sequence that are structurally equivalent to residues in the template sequence.) Real differences between the target structure and template structure in certain regions can lead to alignment errors, as can regions in the alignment where the sequence similarity is simply not high enough to unambiguously identify structurally equivalent residues, even if they do exist. One method to deal with these problems is to vary the parameters used in the alignment. These might include gap penalties, substitution matrix, and environment features of the template. This approach underlies the success behind "metaservers" (Chivian et al., 2005; Fischer, 2003; Kosinski et al., 2003) that combine alignments from different sources (i.e., different individual servers). The implicit assumption behind these methods is that one of them will have used a set of parameters or methods that produce the correct alignment. Various methods of evaluation are then used to identify that alignment.

One approach is to build models based on each alignment and to evaluate these models using a measure of reasonableness that is based on a three-dimensional rather than a one-dimensional score. 3D-Shotgun (Fischer, 2003) builds multiple models and creates a final model from fragments of the initial models using a consensus approach, i.e., the fragments that are assembled into the final model are those that are most often observed (based on a measure of structural similarity) in the initial models. In another approach similar to the fragment-based methods (Kosinski et al., 2003), the model is evaluated using Verify3D (see below), and the alignment in low-scoring regions is sequentially shifted to produce a better scoring model. In the Robetta method (Chivian et al., 2005), a "parametric alignment ensemble" (Jaroszewski et al., 2000) is constructed in which alternate alignments are constructed both by sampling parameters that reflect different measures of similarity and by directly enumerating suboptimal alignments.

While these strategies greatly improve the probability of generating the correct alignment, the problem of recognizing it still remains. Consequently, it is necessary to

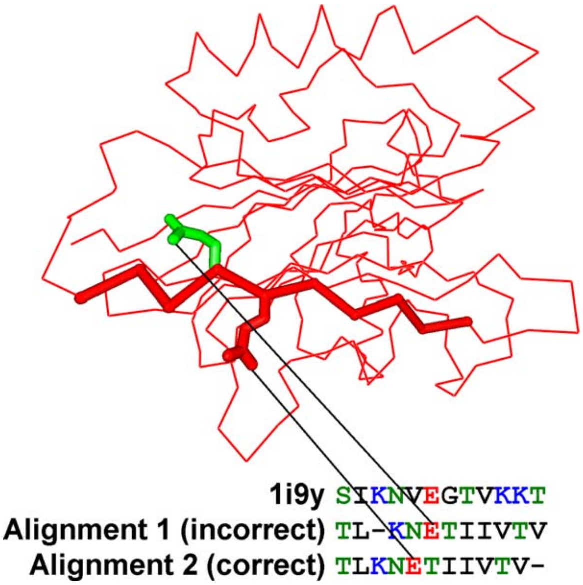
1i9y SIKNVEGTVKKT
Alignment 1 (incorrect) TL-KNETIIVTV
Alignment 2 (correct) TLKNETIIVTV-

Figure 2. Example of an Incorrect Alignment that Can Be Corrected by Three-Dimensional Information

(A) The native structure of bedbug nitrophorin (red wires, PDB code 1ntf).

(B) The sequence of the template and an incorrect (Alignment 1) and a correct (Alignment 2) alignment of the target to the template (synaptojanin, PDB code 1i9y).

build a model using each alignment and evaluate the set of models with some scoring function that reflects the stability of a given conformation. Thus, the one-dimensional alignment problem is ultimately coupled to the problem of building a three-dimensional model and evaluating its conformational energy, at least with respect to models built from other alignments and other possible templates. Figure 2 provides an example taken from CASP5 (Target 142) of how three-dimensional information can be used to evaluate alignments. Despite an unambiguous overall sequence relationship between the target (nitrophorin) and its template (synaptojanin), due to the lack of sequence similarity in the region of the strand shown in Figure 2 (sticks), all predictors at CASP5 who used synaptojanin as a template produced a model in which that strand was misaligned. The incorrect alignment (Alignment 1) shown in Figure 2 produces a model with a glutamate (shown in green) buried on the core side of the strand. Identifying the correct alignment (Alignment 2), where, appropriately, the charged glutamate (shown in red) is exposed, could only be done by making models based on both alignments and evaluating them.

Model Building and Refinement

There are several tools available that can be used to produce a three-dimensional model. The amount of manual intervention required by these programs varies. Swiss-Model (Schwede et al., 2004) allows a user to simply input the sequence of a target and receive a model (assuming a homolog exists in the PDB). NEST (Petrey et al., 2003) requires an already-determined alignment of a target to its template. 3D-Jigsaw allows manual intervention at each stage of the prediction process described above (Bates et al., 2001). MODELLER (Sali and Blundell, 1993) contains a suite of tools for performing each of the stages of a prediction and also has the ability to incorporate information from multiple templates when building the model. The advantages and disadvantages of some of these programs have recently been analyzed (Wallner and Elofsson, 2005).

All of these programs produce a model of the target that is as similar as possible to the template(s) on which it was based. While this is sufficient for many applications, there will be cases in which important features of a model are in regions that are structurally distinct from the template. These will often involve side chains that differ in the template and target and loops between secondary structure elements that can have quite different conformations in the two structures, especially if they have a different length. Both side chain and loop modeling programs typically operate on the assumption that the secondary structure elements of the target structure will be identical to those in the template structure. For the calculation of side chain conformations, the most commonly used methods exploit the observed relationship between side chain conformation and backbone conformation and typically use a "rotamer library" generated from a database of known structures (Dunbrack, 2002). Methods differ in the way in which rotamers are sampled and the energy function used to evaluate the individual conformations (see below). At present, it is possible to predict the conformations of buried side chains with close to experimental accuracy (Canutescu et al., 2003; Xiang and Honig, 2001). Agreement with observed structures is less good for surface side chains, but, in such cases, the crystal structure may place constraints on the side chain conformation that would not be present for a protein free in solution (Eyal et al., 2005; Jacobson et al., 2002).

Loop modeling programs typically build a starting model of the loop in "open" conformation (in which one end of the loop is not connected to its succeeding residue) and then closing the loop using various algorithms (Kolodny et al., 2005a). This process is repeated a number of times using different starting conformations, and resulting conformations are then evaluated using some energy function (see below). Remarkably accurate predictions are now available for loop lengths of up to nine (less than 1 Å backbone atom rmsd from the native structure), although achieving results of such accuracy involves heavy computational demands and requires that the crystal environment be taken into account. Overall, it appears that a combination of extensive sampling and a conformational energy evaluation with a sufficiently detailed energy function can produce highly accurate results (Jacobson et al., 2004; Looger et al., 2003). This is reassuring in the sense that the current limitations of side chain and loop modeling appear to be computational rather than conceptual.

A fundamental problem, even in the most straightforward cases of homology modeling, is that there are generally real differences between the target and template structures that will reduce the accuracy of the model. These differences can involve different relative orientations of secondary structure elements, different numbers of secondary structure elements, and large insertions and deletions in different regions of the structure. The refinement problem of beginning with an incorrect model (for example, 2–4 Å rmsd difference between the template and actual target structures) and improving its accuracy has not been solved. Attempting to do this has become an active research area, and success in loop and side chain modeling, admittedly much easier problems, suggests that real progress can be expected, although this may come at significant computational cost. Indeed, some of the most impressive results reported to date can involve as much as 4 days of computer time per protein for a 12 residue loop (Jacobson et al., 2004), or 150 days of computer time for small proteins (Bradley et al., 2005).

Model Evaluation

As opposed to experimental structure evaluation, we do not yet have reliable procedures to assess the quality of a model. There are programs available that determine whether a model satisfies standard steric and geometric criteria (Laskowski et al., 1993; Vriend, 1990). Each of the tools used in the construction of a model, template selection, alignment, model building, and refinement has its own internal measures of quality, but, ultimately, the most meaningful criterion for the quality of a model is its conformational energy. Consequently, some scoring function that reflects this energy must be applied in order to decide between the tens, hundreds, or even thousands of possible models that may be produced in a prediction.

In order to deal with such a large number of possible conformations, a hierarchical approach to model evaluation is often used. In such a procedure, simplified and easy-to-evaluate scoring functions are used to rank all the original models so that a subset can be chosen for more detailed and computationally costly evaluation. A commonly used scoring function is Verify3D (Eisenberg et al., 1997), which evaluates segments of the model based on how well the environments of the residues in that segment (e.g., burial, secondary structure) correlate with their observed propensities for being in those environments. Statistics-based scoring functions, such as ProsaII (Sippl, 1993), derive a measure of the stability of a polypeptide from the frequency that the interactions (atom-atom or residue-residue) seen in that conformation appears in the database of known structures. Functions such as these are easy to evaluate since they depend only on the distance between pairs of atoms. There are many variants of statistics-based potentials (Oldziej et al., 2005; Samudrala and Moult, 1998; Zhang et al., 2004).

More detailed all-atom measures of conformational stability can also be used. These make use of molecular mechanics force fields of the type used in molecular dynamics simulations (Lazaridis and Karplus, 2000) and, to be reliable, should include a proper description of the aqueous solvent. Such methods have recorded impressive successes in their ability to fold protein fragments from unfolded conformations (Pande et al., 2003; Zhu et al., 2005) and have also proved to be successful in applications to the "decoy" problem, the ability to select the experimentally determined X-ray structure from among a large set of alternate conformations of the same polypeptide chain (Fogolari and Tosatto, 2005; Petrey and Honig, 2000). In other words, it is with these functions that one is most likely to identify a native-like model if it is present in the set of models produced for a given target.

The ability to identify the native conformation from a set of decoys offers confidence, as do successes in loop modeling, that our understanding of the conformational stability of proteins is reasonably accurate. The major challenge to the field at the current stage of development is sampling and evaluating enough conformations so that one has a chance of finding the native state. This is not a new problem, and it will not be easy to solve. In principle, one might expect that molecular dynamics methods that can fold protein fragments from disordered states could be applied to an incorrect model that is relatively close to the native structure and refine it to a conformation that is near native. However, this goal has not yet been realized. The ultimate solution will probably require a combination of improved alignment methods, finding structural templates for each problematic region of a structure, and the use of improved scoring functions and sampling procedures.

Biological Applications of Structure Prediction

The previous sections have outlined the modeling process and have emphasized the fact that it is difficult to be certain of the quality of a given model. However, we now have broad experience in the general problem of model evaluation, much of it coming from the CASP experiments. Sali and coworkers have stated that models based on query/template alignment of greater than about 40% have comparable accuracy to NMR structures, based on comparisons to X-ray-determined structures (Sali et al., 1995). However, there are examples in CASP in which good models are built at quite low levels of sequence identity. Without more reliable methods of model evaluation, however, it is, in general, impossible to know whether this level of accuracy has been achieved, and errors such as those described in Figure 2 will invariably crop up. The central question, then, is whether they are close enough to reveal useful biological insights. The answer to this question of course depends on the application, but homology models are, in particular, often quite useful. They are, of course, generally not as accurate as experimentally determined structures, but they often reveal important information and provide a basis for the design and interpretation of experiments in a way that would not be possible in the absence of a structural hypothesis. This section offers a few examples of some recently published applications.

Homology modeling is extensively used in structure-based drug design as discussed in detail in a recent review (Jacobson and Sali, 2004). An important application is the analysis of the structural differences that are responsible for differences in the specificity of ligand binding among members of a protein family. Proteases are implicated in many diseases, and it is important for the design of therapeutic agents to understand the structural basis for their differences in activity and specificity. Recently, Caffrey et al. (2005) used homology model of *Schistosoma japonicum* cathepsin D to identify the structural differences between that protein and its human homolog that were responsible for differential

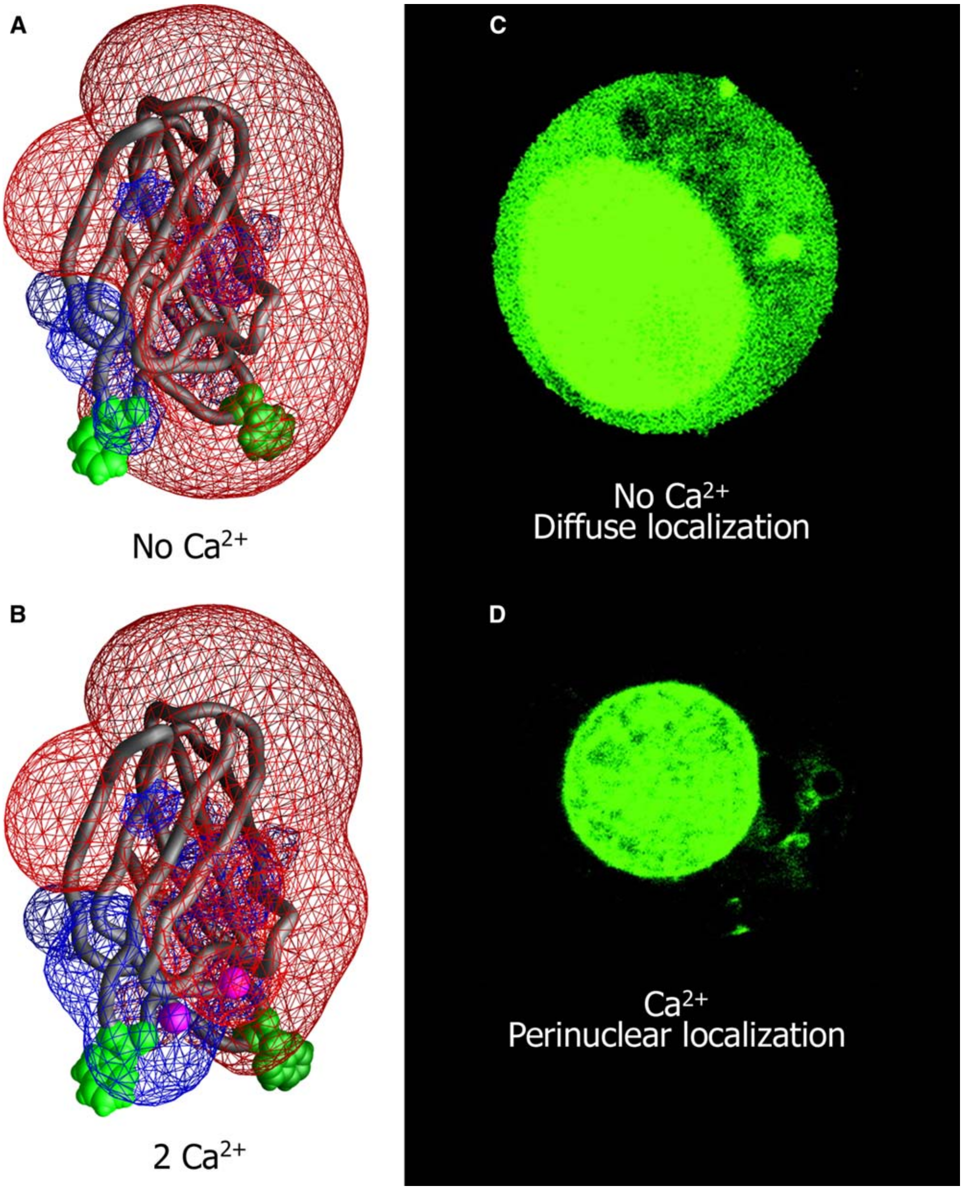

Figure 3. Subcellular Localization of the 5-Lipoxygenase C2 Domain: Models to Biochemistry to Cells

(A and B) Ca trace of the homology model of the C2 domain of 5LO (Kulkarni et al., 2002). The green atoms correspond to some of the hydrophobic residues predicted to be on the calcium binding loops, the magenta spheres in (B) are the bound calcium ions, and the red and blue meshes are the −25 and +25 mV equipotential contours as calculated and imaged in GRASP for 0.1 M KCl.

(C and D) HEK293 cells transiently transfected with the GFP-tagged 5LO C2-like domain were treated with ionomycin. The images were taken before (C) and 5 min after (D) ionomycin treatment. Images provided by Diana Murray and modified from Kulkarni et al. (2002) with permission.

binding of certain types of cathepsin D inhibitors. They used this information to design inhibitors that show greater specificity to the worm version of the protein. Homology models have also been used to propose novel catalytic mechanisms for hydrolysis of the peptide bond in homologs in which active site residues are not conserved (Bjelic and Aqvist, 2004).

Homology models have been used extensively by Murray and coworkers (Yu et al., 2004) to characterize the role of electrostatic interactions in membrane binding properties of peripheral membrane proteins. The models have been used as a basis for predictions that have been central in developing an understanding of the underlying forces that drive membrane binding in different signaling pathways. One example is shown in Figure 3, which contains a model of the C2 domain of 5-lipoxegenase (5LO) and its electrostatic potentials (Kulkarni et al., 2002). In the absence of Ca^{+2}, the C2 domain is predicted to be highly negatively charged (Figure 3A) and thus is unable to bind appreciably to phospholipid vesicles due to electrostatic repulsive forces. In agreement with this prediction, experiments show that, in the absence of Ca^{+2}, the 5LO C2 domain is distributed in the cytosol (Figure 3C). The model also predicts that the C2 domain binds two Ca^{+2} ions that are coordinated by aspartate residues. As shown in Figure 3B, Ca^{+2} binding dramatically reduces the negative potential in this region, suggesting that the C2 domain should bind to neutral lipids mediated by the exposure of hydrophobic residues on the Ca^{+2} binding loops (shown in green); the remaining surrounding negative potential should screen against the negatively charged inner leaflet of the plasma membranes. Consistent with this prediction, the 5LO C2 domain is seen to localize solely to perinuclear membranes that are relatively enriched in the electrically neutral zwitterionic phosphatidylcholine (Figure 3D). Applications such as these provide a glimpse of how homology models can be used effectively when combined with experimental measurements to characterize the properties of proteins of unknown structure.

The surface properties of proteins are important in many other areas of biology (Honig and Nicholls, 1995). Recently, Xu et al. (2005) used homology modeling combined with experiment to identify the structural differences between members of the ErbB family responsible for differential binding to the molecular chaperone Hsp90. In particular, by comparing homology models of ErbB2 (which binds to Hsp90) to the crystal structure of ErbB1 (which does not), they were able to identify the most significant structural difference between the two proteins: a difference in the surface properties of an eight residue loop that they had experimentally determined to be linked to Hsp90 binding. By modeling point mutations, they found that the difference in surface properties was due primarily to the presence/absence of a single aspartate residue and confirmed that this single residue was responsible for the differential binding by making the point mutation.

Although structure prediction in the absence of clear homology is an uncertain enterprise, it can, especially in the hands of experts, be remarkably effective. As an example, Chmiel et al. (2005) were able to identify the catalytic residues of the restriction enzyme NaIIV, a protein that had no detectable sequence homolog, either with or without a structure, in the most commonly used nonredundant databases. Searching farther afield, however, they were able to find a single sequence homolog in the recently collected set of sequences from the Sargasso Sea (Venter et al., 2004). As described above, incorporating some description of the variability of residues in the description of target and template sequences improves the ability to detect remote homologs. Using a sequence alignment of the NaIIV and this newly discovered protein to search the template databases, Chmiel et al. were able to confidently predict that NaIIV had a fold similar to EcoRV with a variety of methods. Using the "Frankenstein monster" approach (Kosinski et al., 2003) to build a model based on EcoRV, they were able to identify the catalytic residues by superimposing the model on the template EcoRV. Their prediction was confirmed by mutating those residues.

Conclusion

Protein structure prediction has been thought of as a "grand challenge" for some time. There has been rapid progress in the past few years, much of it made possible by the massive amounts of data that have become available for analysis. These include, first and foremost, the large number of protein structures that can be used as templates for modeling other proteins, but the explosion of sequence information has also made it possible to identify conservation patterns that can be exploited in structure prediction. This growth in data has led to the development of new structure-based analysis tools that now make it possible to find structural homologs for different regions of a given sequence. There has, in addition, been a parallel development in the ability to predict regions of structures (e.g., loops and side chains) based on physical-chemical methods alone. Moreover, there has been increasing integration of bioinformatics and biophysical techniques, and there is little question that there will be much more of this in the coming years. There is, indeed, great excitement in the field and great expectations. We have tried in this review to give the reader a flavor of what has been accomplished, where the challenges lie, and how structure prediction can be used today.

We have in hand a series of powerful tools that can significantly expand the range of applicability of structural biology. What we have not done is to provide a formula for when a model can be used in a particular application for the simple reason that no reliable formula currently exists. This places a burden on the nonexpert and the expert alike. How are we to deal with an approach that can be so powerful and useful on the one hand, and of such uncertain reliability on the other? Can one use automatic servers or databases of models with any confidence? The best answer to these questions is to view each model as a testable hypothesis whose initial degree of validity can be evaluated by controls on comparable systems or by previous experience with problems of similar complexity. In each case, one should look for regions in the sequence that are different in different alignment methods, and regions that do not score well with standard measures such as Verify-3D and ProsaII. It is our opinion that reports on reliability tests such as these, together with measures of statistical significance for the sequence alignment, should accompany every manuscript that contains a homology model. Further tests might include "control modeling exercises" in which models of other members of proteins of known structure that are homologous to the query protein are built and then evaluated in terms of the extent to which the property or function of interest, which can be extracted from the native structure, is well reproduced in the model.

The application of existing structure and function prediction methods has the potential to have significant impact on many areas of biology, and the impact of these methods will continue to grow, assuming the rapid rate of progress in the development of computational tools retains its current pace. We have not reviewed here the large number of methods that have been developed to predict function from structure, but a number of excellent recent reviews (Rost et al., 2003) and Web sites (Laskowski et al., 2005) are available. Learning the elements of structure prediction may prove to be a valuable investment for researchers interested in applying the insights of three-dimensional protein structure to their research areas. Careful analysis of modeled structures in light of experimental data, while retaining a critical eye to possible shortcomings in the models, can provide potentially dramatic new structural insights about problems that were previously inaccessible to this type of analysis.

Acknowledgments

We are pleased to acknowledge the contributions of our fellow lab members, Marc Fasnacht, Rachel Kolodny, Andy Kuziemko, Peng Liu, Cinque Soto, Chris Tang, Li Xie, and Jiang Zhu, to the collaborative research on the subject areas covered in this review. We are also grateful to Professor Diana Murray for her careful reading of the manuscript and for preparing Figure 3. Support from NIH grants GM-30518 (B.H.) and from the Northeast Structural Genomics Consortium (NESG) GM P50-62412 (PI, Guy Monteleone) is gratefully acknowledged.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Bates, P.A., Kelley, L.A., MacCallum, R.M., and Sternberg, M.J.E. (2001). Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins *45*, 39–46.

Bjelic, S., and Aqvist, J. (2004). Computational prediction of structure, substrate binding mode, mechanism, and rate for a malaria protease with a novel type of active site. Biochemistry *43*, 14521–14528.

Bradley, P., Misura, K.M.S., and Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. Science *309*, 1868–1871.

Bystroff, C., and Baker, D. (1997). Blind predictions of local structure in CASP2 targets uding the I-sites library. Proteins Suppl. *1*, 167–171.

Caffrey, C.R., Placha, L., Barinka, C., Hradrilek, M., Dostal, D., Sajid, M., McKerrow, J.H., Majel, P., Konvalinka, J., and Vondrasek, J. (2005). Homology modeling and SAR analysis of *Shistosoma japonicum* cathepsin D (SjCD) with statin inhibitors identify a unique active site steric barrier with potential for the design of specific inhibitors. Biol. Chem. *386*, 339–349.

Canutescu, A.A., Shelenkov, A.A., and Dunbrack, R.L., Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci. *12*, 2001–2014.

Chivian, D., Kim, D.E., Malmström, L., Schonbrun, L., Rohl, C.A., and Baker, D. (2005). Prediction of CASP-6 structures using automated Robetta protocols. Proteins. Published online September 26, 2005. 10.1002/prot.20733.

Chmiel, A.A., Radlinska, M., Pawlak, S.D., Krowarsch, D., Bujnicki, J.M., and Skowronek, K.J. (2005). A theoretical model of restriction endonuclease NlaIV in complex with DNA, predicted by fold recognition and validated by site-directed mutagenesis and circular dichroism spectroscopy. Protein Eng. Des. Sel. *18*, 181–189.

Dunbrack, J.R.L. (2002). Rotamer libraries in the 21st century. Curr. Opin. Struct. Biol. *12*, 431–440.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755–763.

Eisenberg, D., Luthy, R., and Bowie, J.U. (1997). Verify3D: assessment of protein models with three dimensional profiles. Methods Enzymol. *277*, 396–404.

Eyal, E., Gerzon, S., Potapov, V., Edelman, M., and Sobolev, V. (2005). The limit of accuracy of protein modeling: influence of crystal packing on protein structure. J. Mol. Biol. *351*, 431–442.

Fischer, D. (2003). 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins *51*, 434–441.

Fogolari, F., and Tosatto, S.C.E. (2005). Application of MM/PBSA colony free energy to loop decoy discrimination: toward correlation between energy and root mean square deviation. Protein Sci. *14*, 889–901.

Fox, J.A., Butland, S.L., McMillan, S., Campbell, G., and Ouellette, B.F.F. (2005). The bioinformatics links directory: a compilation of molecular biology web servers. Nucleic Acids Res. *33*, W3–W24.

Friedberg, I., and Godzik, A. (2005). Connecting the protein structure universe by using sparse recurring fragments. Structure *13*, 1213–1224.

Ginalski, K., Grishin, N.V., Godzik, A., and Rychlewski, L. (2005). Practical lessons from protein structure prediction. Nucleic Acids Res. *33*, 1874–1891.

Honig, B., and Nicholls, A. (1995). Classical electrostatics in biology and chemistry. Science *268*, 1144–1149.

Jacobson, M., and Sali, A. (2004). Comparative Protein Structure Modeling and its Applications to Drug Discovery. In Annual Reports in Medicinal Chemistry, J. Overington, ed. (London: Academic Press), pp. 259–276.

Jacobson, M.P., Friesner, R.A., Xiang, Z., and Honig, B. (2002). On the role of the crystal environment in determining protein side-chain conformations. J. Mol. Biol. *320*, 597–608.

Jacobson, M.P., Pincus, D.L., Rapp, C.S., Day, T.J.F., Honig, B., Shaw, D.E., and Friesner, R.A. (2004). A hierarchical approach to all-atom protein loop prediction. Proteins *55*, 351–367.

Jaroszewski, L., Rychlewski, L., and Godzik, A. (2000). Improving the quality of twilight-zone alignments. Protein Sci. *9*, 1487–1496.

Kelley, L.A., MacCallum, R.M., and Sternberg, M.J.E. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. J. Mol. Biol. *299*, 501–522.

Kihara, D., and Skolnick, J. (2003). The PDB is a covering set of small protein structures. J. Mol. Biol. *334*, 793–802.

Kolodny, R., Guibas, L., Levitt, M., and Koehl, P. (2005a). Inverse kinematics in biology: the protein loop closure problem. Int. J. Robot. Res. *24*, 151–163.

Kolodny, R., Koehl, P., and Levitt, M. (2005b). Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J. Mol. Biol. *346*, 1173–1188.

Kosinski, J., Cymerman, I.A., Feder, M., Kurowski, M.A., Sasin, J.M., and Bujnicki, J.M. (2003). A "Frankenstein's monster" approach to comparative modeling: merging the finest fragments of fold-recognition models and iterative model refinement aided by 3D structure evaluation. Proteins Suppl. *53*, 369–379.

Kulkarni, S., Das, S., Funk, C.D., Murray, D., and Cho, W. (2002). Molecular basis of the specific subcellular localization of the C2-like domain of 5-lipoxygenase. J. Biol. Chem. *277*, 13167–13174.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. *26*, 283–291.

Laskowski, R.A., Watson, J.D., and Thornton, J.M. (2005). ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res. *33*, W89–W93.

Lazaridis, T., and Karplus, M. (2000). Effective energy functions for protein structure prediction. Curr. Opin. Struct. Biol. *10*, 139–145.

Lo Conte, L., Ailey, B., Hubbard, T.J.P., Brenner, S.E., Murzin, A.G., and Chothia, C. (2000). SCOP: a structural classification of proteins database. Nucleic Acids Res. *28*, 257–259.

Looger, L.L., Dwyer, M.A., Smith, J.J., and Hellinga, H.W. (2003). Computational design of receptor and sensor proteins with novel functions. Nature *423*, 185–190.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. *29*, 291–325.

Marti-Renom, M.A., Madhusudhan, M.S., and Sali, A. (2004). Alignment of protein sequences by their profiles. Protein Sci. *13*, 1071–1087.

McGuffin, L.J., and Jones, D.T. (2003). Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics *19*, 874–881.

Moult, J., Fidelis, K., Zemla, A., and Hubbard, T. (2003). Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins Suppl. *53*, 334–339.

Murray, P.S., Li, Z., Wang, J., Tang, C.L., Honig, B., and Murray, D. (2005). Retroviral matrix domains share electrostatic homology: models for membrane binding function throughout the viral life cycle. Structure *13*, 1521–1531.

Nanias, M., Chinchio, M., Oldziej, S., Czaplewski, C., and Scheraga, H.A. (2005). Protein structure prediction with the UNRES force-field using Replica-Exchange Monte Carlo-with-Minimization; comparison with MCM, CSA, and CFMC. J. Comput. Chem. *26*, 1472–1486.

Oldziej, S., Czaplewski, C., Liwo, A., Chinchio, M., Nanias, M., Vila, J.A., Khalili, M., Arnautova, Y.A., Jagielska, A., Makowski, M., et al. (2005). Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. Proc. Natl. Acad. Sci. USA *102*, 7547–7552.

Pande, V.S., Baker, I., Chapman, J., Elmer, S.P., Khaliq, S.P., Larson, S.M., Rhee, Y.M., Shirts, M.R., Snow, C.D., Sorin, E.J., and Zagrovic, E.J. (2003). Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. Biopolymers *68*, 91–109.

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., et al. (2005). The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res. *33*, D247–D251.

Pearson, W.R., and Sierk, M.L. (2005). The limits of protein sequence comparison? Curr. Opin. Struct. Biol. *15*, 254–260.

Petrey, D., and Honig, B. (2000). Free energy determinants of tertiary structure and the evaluation of protein models. Protein Sci. *9*, 2181–2191.

Petrey, D., Xiang, Z., Tang, C.L., Xie, L., Gimpelev, L., Mitros, T., Soto, C.S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., et al. (2003). Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. Proteins Suppl. *53*, 430–435.

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O., and Ofran, Y. (2003). Automatic prediction of protein function. Cell. Mol. Life Sci. *60*, 2637–2650.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779–815.

Sali, A., Potterton, L., Yuan, F., van Vlijmen, H., and Karplus, M. (1995). Evaluation of comparative protein modeling by modeller. Proteins *23*, 318–326.

Samudrala, R., and Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol. *275*, 895–916.

Sanchez, R., and Sali, A. (1997). Advances in comparative protein-structure modelling. Curr. Opin. Struct. Biol. *7*, 206–214.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C. (2004). SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. *31*, 3381–3385.

Shindyalov, I.N., and Bourne, P.E. (2000). An alternative view of protein fold space. Proteins *38*, 247–260.

Simons, K.T., Bonneau, R., Ruczinski, I., and Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins Suppl. *37*, 171–176.

Sippl, M.J. (1993). Recognition of errors in three-dimensional structures. Proteins *17*, 355–362.

Szustakowski, J.D., Kasif, S., and Weng, Z. (2005). Less is more: towards an optimal universal description of protein folds. Bioinformatics *21*, ii66–ii71.

Tang, C.L., Xie, L., Koh, I.Y.Y., Posy, S., Alexov, E., and Honig, B. (2003). On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. J. Mol. Biol. *334*, 1043–1062.

Venclovas, C. (2003). Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance. Proteins Suppl. *53*, 380–388.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science *304*, 66–74.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. *8*, 52–56.

Wallner, B., and Elofsson, A. (2005). All are not equal: a benchmark of different homology modeling programs. Protein Sci. *14*, 1315–1327.

Xiang, Z., and Honig, B. (2001). Extending the accuracy limits of prediction for side-chain conformations. J. Mol. Biol. *311*, 421–430.

Xu, W., Yuan, X., Xiang, Z., Mimnaugh, E., Marcu, M., and Neckers, L. (2005). Surface charge and hydrophobicity determine ErbB2 binding to the Hsp90 chaperone complex. Nat. Struct. Mol. Biol. *12*, 120–126.

Yang, A.-S., and Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. J. Mol. Biol. *301*, 691–711.

Yu, J.W., Mendrola, J.M., Audhya, A., Singh, S., Keleti, D., DeWald, D.B., Murray, D., Emr, S.D., and Lemmon, M.A. (2004). Genome-wide analysis of membrane targeting by *S. cerevisiae* Pleckstrin homology domains. Mol. Cell *13*, 677–688.

Zhang, Y., and Skolnick, J. (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. Proc. Natl. Acad. Sci. USA *101*, 7594–7599.

Zhang, Y., and Skolnick, J. (2005). The protein structure prediction problem could be solved using the current PDB library. Proc. Natl. Acad. Sci. USA *102*, 1029–1034.

Zhang, C., Liu, S., Zhou, H., and Zhou, Y. (2004). An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Sci. *13*, 400–411.

Zhou, H., and Zhou, Y. (2004). Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins. *55*, 1005–1013.

Zhu, J., Alexov, E., and Honig, B. (2005). Comparative study of generalized born models: born radii and peptide folding. J. Phys. Chem. B *109*, 3008–3022.