Protein structure prediction
Bissan Al-Lazikani, Joon Jung, Zhexin Xiang and Barry Honig*

The prediction of protein structure, based primarily on sequence and structure homology, has become an increasingly important activity. Homology models have become more accurate and their range of applicability has increased. Progress has come, in part, from the flood of sequence and structure information that has appeared over the past few years, and also from improvements in analysis tools. These include profile methods for sequence searches, the use of three-dimensional structure information in sequence alignment and new homology modeling tools, specifically in the prediction of loop and side-chain conformations. There have also been important advances in understanding the physical chemical basis of protein stability and the corresponding use of physical chemical potential functions to identify correctly folded from incorrectly folded protein conformations.

**Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D-PSSM</td>
<td>three-dimensional position specific scoring matrix</td>
</tr>
<tr>
<td>CASP</td>
<td>critical assessment of structure prediction</td>
</tr>
<tr>
<td>HMM</td>
<td>hidden Markov model</td>
</tr>
<tr>
<td>PSI-BLAST</td>
<td>position-specific iterative basic local search and alignment tool</td>
</tr>
<tr>
<td>msd</td>
<td>root mean square deviation</td>
</tr>
<tr>
<td>SCOP</td>
<td>structural classification of proteins</td>
</tr>
</tbody>
</table>

**Introduction**

The prediction of the three-dimensional structure of a protein when only the amino-acid sequence is known has been a problem of major interest for many years. Approaches have ranged from purely ab-initio methods that are based entirely on physical chemical principles, to homology methods that are based primarily on the information available in sequence and structural databases. Threading, or fold-recognition methods lie between these two extremes and involve the identification of a structural template that most closely resembles the structure of a query sequence. These three approaches define the categories of the notorious CASP (critical assessment of structure prediction; [1**]) ‘experiment’, which is run every two years. CASP is made possible by crystallographers and NMR spectroscopists who submit sequences whose structures are about to be determined to the CASP organizers. These sequences are made available over the web to CASP predictors who very often invest a tension-filled summer in the problem of structure prediction. (Some of the authors of this article are currently recovering from such a summer.) The ‘contestants’ and other interested spectators then assemble at Asilomar, California in December, where a chosen judge, one for each category, assesses predictions. The very existence of CASP is a testimony to the fact that protein structure prediction has become a very real and serious enterprise. The results of the previous experiments have been published in special issues of *Proteins: Structure, Function, Genetics* [2]. They are required reading for people interested in this area.

One of the interesting lessons that can be learned from the progression of the CASP experiments is that the overlap between the three categories has increased dramatically. Overall, the past few years have witnessed considerable progress in all aspects of protein-fold prediction. Advances in simulation methodology and in the understanding of the forces that drive protein folding have had a major impact on ab-initio methods, whereas rapidly expanding databases and new sequence and structure analysis algorithms have dramatically increased the accuracy of homology models. In parallel, there has been an increasing interplay between the two approaches as physical chemical methods have been used to refine homology models while information derived from databases has been incorporated into ab-initio schemes.

The goals of protein-structure prediction are varied and range from the intellectual challenge of solving a clearly fascinating puzzle to the diverse applications that become possible once accurate prediction becomes a reality. Until recently, with few exceptions, the prediction of protein structure has been of greater conceptual than practical importance in that predictions were rarely accurate enough to be used, for example, to deduce biological function or to facilitate the structure-based design of new pharmaceuticals. This situation has changed dramatically as homology models have become increasingly accurate and as the nature and range of their application has changed because of other developments. For example, the increased use of combinatorial libraries in lead-compound discovery has reduced the level of accuracy needed from models, because the overall features of the binding site can be extremely useful in biasing libraries towards compounds that have features that are compatible with a given binding site. From a biological perspective, homology models can be very useful in deducing the function of proteins of undetermined structure and in generalizing and extending patterns observed from the structure of a small number of proteins to a large protein family.

Homology modeling has recently assumed an increased importance with the advent of structural genomics initiatives around the world. One of the frequently stated goals of high-throughput experimental structure determination is to obtain enough protein structures so that the rest can
be reliably predicted using homology methods. Although this is clearly a long-range goal, it is quite likely that homology modeling will assume an increasingly important role in both biological and chemical applications. Fold recognition has also gained in stature and has become an important tool that supplements sequence-based methods to detect remote homologs. *Ab-initio* structure methods achieved a number of remarkable successes at CASP III and are extremely important, not only for what they can accomplish but for what they can teach us about protein folding. This review focuses primarily on homology modeling, although lessons learned from other categories of structure prediction will be considered as well. In addition, its emphasis is more on the physical chemical rather than statistical description of protein energetics; however, information derived from database analysis will also be discussed, particularly with regard to sequence alignment.

In addition to the collection of articles in the CASP III supplement [2], an excellent review by Sali and co-workers [3••] provides a comprehensive discussion of the issues involved in homology modeling. The process begins with the choice of the most appropriate template for a given sequence of unknown structure. In general, more than one template may be required if, for example, one part of a sequence fits best to one template and one part to another. In addition, it may be beneficial to use multiple templates simultaneously, for example by averaging over all of them, if it is not clear which is most appropriate. Once a template is chosen, and often in parallel to this process, a good alignment of the query sequence and the template needs to be generated. The next step is to construct a model based on the alignment. Some parts of the model may be obtained directly from the known structure(s) but for other parts there may be no appropriate template in the database. These regions often correspond to surface loops but, in addition, many side chains will in general be different to those of the template structure. There has been significant recent progress on many of these problems, as will be discussed below.

**Detecting homology and obtaining alignments: sequence profiles and iterative searches**

The past few years have witnessed the emergence of sequence profile methods as the optimal approach in homology detection. This has resulted in considerable increases in sensitivity over more traditional pairwise alignment methods. Position-specific profile search methods such as PSI-BLAST [4] and hidden Markov models (HMMs) [5], as implemented in the SAM [6] and HMMER [7] packages, have vastly improved the accuracy of sequence alignments and have extended the boundaries of detectable sequence similarity. Although a major goal of this effort has been remote homolog detection, an important side benefit has been significant improvement in alignment quality, even at levels of sequence identity for which pairwise alignment methods are known not to work. This, in turn, has had a positive impact on the starting alignments used in homology modeling. Indeed, at least part of recent improvements in homology modeling can be traced directly to improvement in sequence alignment algorithms.

Structural information has proved invaluable in benchmarking purely sequence-based methods. In particular, because structure is conserved more extensively than sequence, it has become commonplace to test the sensitivity of sequence alignments through their ability to detect relationships between proteins that are evident only from structure. The SCOP (structural classification of proteins) database [8] has been used widely for this purpose because it groups proteins according to evolutionary relationships as determined by structural and functional similarity. Using SCOP, Müller et al. [9] extended the coverage of PSI-BLAST simply by demanding that the query sequence needs only to match a single homolog in the target database. By analyzing a number of fold families whose relationships were only detectable by structural comparison, Aravind and Koonin [10•] demonstrated the increased sensitivity in detecting distant homologs that can be obtained by PSI-BLAST through careful selection of queries and searching in a multi-step process. Using SCOP as a benchmark to compare different sequence search methods, Lindahl and Elofsson [11•] observed that different methods cover overlapping, yet different areas of sequence space. Profile-profile comparison methods have begun to appear as opposed to the profile-against-sequence method of PSI-BLAST. For example, Godzik and co-workers [12] have reported two such methods and reported improved sensitivity relative to PSI-BLAST.

Structural information can also be used in other ways. Quantitative measures of structural distance can be used to produce precise relationships between sequence and structural similarity [13•]. Moreover, multiple structure alignments can be used to generate structure-based sequence profiles that are generally superior to purely sequence-based profiles [14•]. The three-dimensional position specific scoring matrix, 3D-PSSM [15••], uses sequence and structure alignments as well as secondary structure and solvent accessibility to construct descriptive position-specific matrices for each domain in a non-redundant structural database. Panchenko et al. [16•] have reported a method that combines PSI-BLAST sequence profiles with residue contact potentials to improve the detection of remote homologs. GenTHREADER [17••], which was designed with genomic-scale automated prediction in mind, combines the tools of profile sequence alignments and threading to improve the speed and performance of fold-recognition. These methods can clearly be used to improve alignments for homology modeling and thus have the potential to extend the applicability of homology modeling to increasingly lower levels of sequence similarity.

**Building homology models**

As discussed in the review of Sali and co-workers [3••], a number of methods, and programs, are available to build homology models once an alignment has been determined. We will not review the problem of model construction in this
paper. We will, however, summarize recent progress in loop modeling and in the prediction of side-chain conformations.

**Loop prediction**

The basic goal is to predict the conformation of a loop that is fixed at both ends by the protein backbone. *Ab-initio* methods of loop prediction involve the generation of a large number of randomly chosen candidate conformations and their evaluation with energetic or other criteria. Database methods generate trial conformations based either on sequence relationships to loops of known structure, or based on geometric criteria such as the distance between the amino and carboxyl termini of the loop in question. Once loops are generated in this way, energetic criteria are often applied to select the most likely candidate. Recent papers on loop prediction include those based on *ab-initio* methods [18••,19,20••], database related methods [21,22] or a combination of both [23,24].

It is obviously important that near-native conformations be present among the trial conformations generated in the first step of loop modeling. Adequate sampling does not appear to be a problem if a large enough number of loops is generated randomly; indeed, Rapp and Friesner [20••] were able to generate near-native conformations for even a 12-residue loop. However, database methods generate a much smaller number of trial conformations and the lack of a large enough template library to cover the many possible conformations of longer loops (more than five residues not including stem, or anchoring, residues that are kept fixed) limits their utility for these cases [23]. A recent paper on sequence-dependent loop prediction using a database method achieved an average accuracy of only 3.8 Å rmsd for the backbone atoms of eight-residue loops [22]. Vlijmen and Karplus [24] used CHARMM to optimize initial conformations that were selected from the protein database. They report improved results for longer loops but their optimization procedure, which involves simulated annealing, effectively extends the conformation space beyond that provided by the initial conformations. In this sense, their approach is closer to *ab-initio* loop generation.

Because conformational sampling does not appear to be a problem for loops, the quality of the scoring function used to evaluate loop conformations is the major determinant of loop-prediction accuracy. Loop accuracy is usually evaluated in terms of local r.m.s.d (involving the optimal superposition of the predicted and native loop independent of the rest of the structure) or global r.m.s.d (where the r.m.s.d is evaluated with the loop stems kept in place). We prefer the latter measure because the former allows for two loops to be seen as similar, and to have a small r.m.s.d, even if they have very different orientations in the context of the native structure. Rapp and Friesner [20••] used the generalized Born solvation model and the AMBER94 force field to obtain low r.m.s.d values for the two loops they studied. However, their approach still needs to be tested on a larger sample size. Fiser et al. [18••] have recently published an extensive *ab-initio* study on a data set of 40 loops and also report low r.m.s.d from known structures. Using global r.m.s.d as a criterion, Fiser et al. [18••] reported an accuracy of 1.16 Å local r.m.s.d that, based on the conversion factor of 1.5 provided by the authors, translates into less than 2 Å global r.m.s.d for eight-residue loops.

The study of Fiser et al. [18••] utilized a scoring function that included the CHARMM22 force field and statistical preferences taken from protein databases. Scoring functions based entirely on physical chemical potentials and an accurate solvation model have the potential of identifying the native conformation as lowest in energy, but there are cases where lower energy conformations appear [25]. One problem may be that essentially every approach seeks the lowest energy conformation, thus ignoring conformational entropy effects that will favor broad energy wells. We have recently implemented a procedure that takes the shape of the energy well into account and yields highly accurate loop prediction (e.g. 1.6 Å global r.m.s.d for eight-residue loops [Z Xiang and B Honig, unpublished data]).

Most tests of proposed procedures involve generating structures for which the answer is already known and comparing these with known structures. That is, the backbone conformation of anchoring residues identical to that of the native conformation. This does not properly simulate real modeling conditions under which the backbone of the query protein may not be identical to that of the template. In general, loop prediction accuracy degrades as the constraints provided by the loop ends are less accurately defined [18••,26•].

**Side-chain prediction**

The greatest success in the prediction of side-chain conformations has been achieved for core residues where packing constraints significantly simplify the problem. Even for core residues, the accuracy of side-chain prediction degrades when the structure of the backbone is itself not known to a high degree of accuracy. Many side-chain programs are based on rotamer libraries [27], which are generally defined in terms of side-chain torsional angles for preferred conformations of a particular side chain. The resolution of rotamer libraries has increased over time (see, for example, Dunbrack and Karplus [28]) and rotamer libraries have been compiled simply by sampling all angles at some given level of resolution (e.g. Maeyer et al. [29]). As the number of rotamers increases, however, so does the problem of sampling all possible conformations. There have been a variety of approaches developed to deal with the combinatorial problem in side-chain prediction (see, for example [30–34]).

In recent papers, accuracies of about 1 Å r.m.s.d have been reported for core residues in known structures where the backbone has been fixed in the native conformation [34–37]. A number of recent studies suggest that improvements on these values may still be possible. Mendes et al. [38] found, for example, that the use of an intrinsic torsional potential can improve prediction accuracy. Lovell et al. [39] recently reported a novel rotamer library in which internal clashes
between side chain and backbone are removed. This library could, in principle, be used to improve prediction accuracy. We have recently shown (Z. Xiang, B. Honig, unpublished data) that using a very detailed rotamer library, which is based on rotamers that use Cartesian coordinates taken from known structures rather than idealized bond lengths and angles, yields rmsd values relative to the native of only 0.62 Å for core residues. This appears to constitute a significant improvement over existing procedures and demonstrates that the combinatorial problem, usually assumed to confound side-chain prediction, may in fact be of little consequence.

As is the case for loop prediction, side-chain prediction accuracy depends sensitively on the accuracy to which the backbone conformation is known [40]. This suggests the possibility of developing procedures where side-chain and backbone conformations are alternately optimized in an iterative fashion.

**Energetic discrimination**

As is evident from the discussion above, the development of an energy function that is capable of identifying the most stable conformation of a protein, or of some part of a protein, would have enormous impact on the ability to predict protein conformation. Scoring functions used for the evaluation of protein models generally fall into two broad categories (see [41] for a recent discussion). ‘Statistical’ effective energy functions, termed SEEFs by Lazaridis and Karplus [41] are based on the observed properties of amino acids in known structures, and have been widely used in fold recognition and homology modeling applications. Physical effective energy functions, or PEEFs, are based on a direct evaluation of the conformational free energy of a protein. Recent work has demonstrated that such a direct evaluation of the conformational free energy can be at least as successful as statistically based scoring functions (see, for example [42,43]) in distinguishing the native structure of a protein from an incorrectly folded decoy, although generally at greater computational cost [44,45,46*]. A distinct advantage of such physically derived functions is that they are based on well-defined physical interactions, thus making it easier to learn and to gain insight from their performance. Moreover, the success in CASPIII of *ab-initio* methods based on purely physical chemical methods [48**] suggests that our understanding of the forces that drive protein stability may have reached the point where it can be translated into widely applicable computational tools.

One of the major drawbacks of accurate physical chemical description of the folding free energy of a protein is that the treatment of solvation required usually comes at a significant computational expense. Fast solvation models such as the Generalized Born [49] and a variety of simplified scoring schemes [47*,50] may prove to be extremely useful in this regard.

**Conclusions**

For many years, protein-structure prediction was viewed as an important but distant goal. This perspective has changed dramatically with the recent explosion of sequence and structural information and because of computational advances in many different areas. These include pure sequence analysis, structure-based sequence analysis, the conformational analysis of proteins and the understanding of the energetic determinants of protein stability. Evidence for the progress that has been made is the fact that homology modeling has become a widely used tool while fold recognition has been shown to extend the limits of detection of sequence search methods. Large databases of homology models have been constructed [51,52] and the number of such models and their application in different areas is growing at a rapid rate.

Despite this progress, much remains to be done. Sequence alignment is still a serious problem when low levels of sequence identity are involved, and despite recent progress in *ab-initio* methods for loop and side-chain prediction these problems are by no means solved. An additional problem that still plagues structure prediction by homology is that the query structure may differ significantly from the closest available template. There were cases in CASPIII where a predicted model was incorrect in some region simply because the template and target structures were different; for example, a helical segment in one protein that was unwound in the other. Thus, an important goal for future research is to be able to ‘relax’ a structure, for example, with some simulation method, from that of a template (which may just be a good starting guess) to that of the actual sequence in question. Future progress will depend on advances in both computational chemistry and computational biology and the integration of these two fields.

**Acknowledgements**

This work was supported by grant GM30518 from the National Institutes of Health. We thank Donald Petrey, Felix Sheinerman and Lei Xie for their advice. We are grateful to Peter Chen and Elizabeth Valentine for their much appreciated help and involvement in our structure-prediction efforts.

**References and recommended reading**

Papers of particular interest, published within the annual period of review, have been highlighted as:

* of special interest
** of outstanding interest


This paper compares a number of sequence search methods and draws attention to the fact that different methods cover different areas of sequence classification of proteins database for the investigation of sequences and structures.


This paper emphasizes the importance of sequence profile quality and careful selection of query sequences in iterative database searching.


This paper compares a number of sequence search methods and draws attention to the fact that different methods cover different areas of sequence space. Thus using multiple methods would improve coverage.


This paper demonstrates the existence of a relationship between sequence and structural similarity, even for very low levels of sequence identity.


Multiple structure alignments are used to identify conserved sequence and structure features in different protein families.


The use of structurally derived information as well as PSI-BLAST-derived profiles makes 3D-PSSM an intelligent and promising tool for structure prediction.


A comprehensive description of the method that proved highly successful in CASP III.


A nice demonstration of how sequence information aids the performance of threading.


CHARMM22 force field and database-derived statistical potential are used to discriminate the energetics of loop conformations.


An extensive conformational search, a well-tested force field and the Generalized Born solvation model are used to predict the conformation of an 8- and a 12-residue loop in ribonuclease A.


This paper demonstrates the influence of loop stems on loop modeling.


A physical chemical potential function and a novel solvation model successfully discriminate native structures from misfolded decoys.


CHARMM22 force field and a Poisson-Boltzmann solvation model are used to distinguish native structures. It is shown that the Coulomb energy generally favors the native conformation and, on this basis, a simplified scoring function is developed.


