Computational Methods for Protein Folding: Advances in Chemical Physics, Volume 120.
Edited by Richard A. Friesner. Series Editors: I. Prigogine and Stuart A. Rice.

ISBNs: 0-471-20955-4 (Hardback); 0-471-22442-1 (Electronic)

KNOWLEDGE-BASED PREDICTION OF PROTEIN TERTIARY STRUCTURE

PIERRE-JEAN L'HEUREUX, BENOIT CROMP, AND ÉRIC MARTINEAU

Département de Chimie, Université de Montréal, Montréal, Québec, Canada; Centre de Recherche en Calcul Appliqué, Montréal, Québec, Canada; and Protein Engineering Network of Centers of Excellence, Edmonton, Alberta, Canada

JOHN R. GUNN

Schrödinger, Inc., New York, NY, U.S.A.; Centre de Recherche en Calcul Appliqué, Montréal, Québec, Canada; and Protein Engineering Network of Centers of Excellence, Edmonton, Alberta, Canada

CONTENTS

I. INTRODUCTION

Under the general heading of ''protein folding'' there is an ever-increasing body of methodology that has been rapidly evolving over the past few years. The simply stated objective of computationally determining the three-dimensional atomic coordinates of a protein starting from knowledge of the amino acid sequence remains a somewhat idealistic academic challenge, but it has led to the development of a technology base that is gaining in practical applicability. This corresponds to some extent to a shift in philosophy in which a fundamental understanding of the folding process is of less immediate interest than obtaining the best model possible with whatever means are available. Fundamental questions are of course still important and are being actively pursued [1–5], but the field is being driven more and more by the pragmatic approach [6,7]. This is highlighted by the effort being devoted to the CASP experiments, where the emphasis is placed squarely on the bottom line [8]. In this context, the methods used must be tailored to the particular problem at hand, and no available information can be left unused. Much work therefore has been devoted to making use of prior information and accumulated knowledge in the generation

of computer models of proteins. This review will describe some of the ways in which such methods are being incorporated within the traditional *ab initio* framework.

A. The Knowledge-Based Approach

The label of ''knowledge-based'' is to some extent artificial, in that there is a spectrum of methodologies and it is not always easy to draw a clear distinction. The intended contrast is with a purist's *ab initio* approach in which one seeks a numerical solution to the fundamental laws of physics (as one would like to do in quantum chemistry) with no theoretical limit on the problems that can be addressed. A knowledge-based method, on the other hand, requires some form of *a priori* knowledge and is therefore limited in its applicability by the data that are available. If the term is used in its broadest sense, referring to methods that make explicit use of the Protein Data Bank (PDB) of known structures, this would still cover a range extending from methods which require there to be a similar structure in the PDB to those that apply observed patterns in a more general way. In principle, this includes virtually all methods because even the most determined *ab initio* practitioner still has recourse to an empirical force field that typically uses the PDB in its parameterization [9]. Even though such force fields are as general as possible, the reliance on the PDB does represent a real limitation, as anyone who has ever tried to use one to fold a membrane protein can attest.

In the context of the CASP experiments [8], the distinction is drawn between *ab initio* and ''fold recognition'' predictions, but there as well some overlap occurs [7]. Fold recognition often involves some refinement to model parts of the structure not found by homology, and conversely many *ab initio* methods make some use of structural fragments from the PDB. It is precisely this middle ground where the different categories are converging that is of interest and where much recent success can be found. It has become clear that there is a great deal of information to be had in the PDB and that progress is being made by extending the ways in which it can be used. The knowledge-based approach is therefore to develop methods to take advantage of what is there, even if the underlying physical principles are not fully understood.

B. Recent Trends

One of the patterns that has emerged from the CASP experiments is the relative success of the fold-recognition methods in identifying distant homologies, even in some cases where none was originally thought to exist [8]. Until recently, *ab initio* methods lagged far behind, but significant progress is now being made [7]. As mentioned above, however, this is coming from knowledge-based methods that have incorporated some of the methods that have proven successful in comparative modeling and fold recognition. It has been shown that so-called

"hybrid" methods can outperform more traditional fold-recognition and *ab initio* techniques [10]. The more general methods of fold-recognition have also been shown to outperform direct homology modeling in cases of weak homology [11], suggesting that a flexible approach has the potential to cover a broad spectrum of possible targets.

Another pattern that is emerging is an increased recognition that the PDB can be used to identify structural motifs at different scales, not just individual residues (as used to derive contact potentials) or entire domains (as used in traditional fold recognition). Much recent work has gone into using the PDB to develop databases of smaller fragments which can be used to construct protein models [12], and an approach based on using local homology with a fragment library has been shown to be quite successful at generating new folds [13]. This building-block approach has also been used to generate improved sensitivity and more accurate alignments when applied to fold recognition [14].

The trend toward a more generalized approach is also reflected in recent work on scoring functions. It has been shown that traditional empirical potentials perform poorly at discriminating the correct structure [15] and that the functional form of pairwise contact energies is not even sufficient in principle [16]. The importance of evolutionary relationships has also been established, and information from multiple sequences can be used to improve recognition of misfolded structures [17]. This idea has led to the use of conformational tendencies and contact predictions from multiple sequence alignments [18] and the development of scoring functions which take into account sequence homology [19]. Scoring functions can therefore be constructed as a set of complementary components: contributions that are unique to a given sequence, those that depend on a family or class of sequences, and those that apply to all proteins.

C. Practical Considerations

The bottom line in structure prediction is to provide a useful answer to a question that is actually being posed. *Ab initio* predictions alone are rarely accurate enough to be useful; however, as NMR spectroscopy is being used to obtain structures for larger and larger proteins, there is a great practical benefit in using computational methods to aid in this process. Structure prediction methods, when coupled with experimental data, can be used to obtain higher-quality structures [20] and even to help in interpreting and assigning the spectra [21]. For this reason there is a great interest in developing methods that can make the best use of various types of experimental data (often in the form of constraints) in addition to that gleaned from the PDB.

The enormous progress that has been made in genome sequencing has also led to increased efforts in functional genomics; that is, it has enabled the use of prediction techniques to assign probable functions to newly discovered

sequences [22]. In this case, the emphasis is less on obtaining accurate coordinates and more on being able to detect weak homologies in distantly related families of structures. Improved prediction methods therefore have an important role in improving the sensitivity of fold-recognition techniques, providing better alignments, and ultimately allowing weaker relationships to be detected thereby classifying more of the genome. Even though the protein folding problem may still be a long way from being fully solved, there is a great opportunity for knowledge-based methods to have a significant impact in improving structure prediction's bottom line.

II. PROTEIN MODELING

The most direct approach to modeling protein folding would be to carry out a simulation that replicates the actual folding process as it occurs in nature. Although some progress has been made in pursuing that approach [23–25], it remains impractical in most cases for two reasons: The time scale of the folding transition for moderately sized proteins exceeds that which can be attained in simulations, and the physical forces involved are not modeled with sufficient accuracy to ensure the desired outcome. Because highly simplified models are unsuitable for predicting structural details, a different point of view is needed to carry out tractable simulations of realistic models. If one is not interested in the thermodynamics of folding and wishes only to produce the folded structure, any number of nonphysical buildup or pattern-generation techniques could be imagined; however, many methods retain the basic model of a molecular simulation, albeit with a number of simplifying approximations.

A. The Computational Model

The principal simulation paradigm is based on the thermodynamic hypothesis, namely that the equilibrium structure corresponds to the global minimum of the thermodynamic free energy. Whether or not this is strictly true for a given sequence is not known; however, for the purposes of the simulation it is generally assumed that some sort of energy-like function can in principle be constructed for which the native structure is a minimum. This can be thought of as some sort of modified free energy or as a purely empirical scoring function; either way the mathematical problem is the same, namely to find the global minimum. The general problem of global minimization is nontrivially difficult, and therefore additional approximations are required in order to obtain a solution in a reasonable time. The thermodynamic analogy is often used to model this as an annealing process; however, in general any minimization method can be applied.

In its general formulation, a simulation within the framework of global function minimization consists of three basic elements. As mentioned above, the

target function of the minimization must be defined so as to allow comparison of different possible structures. Secondly, there must be a procedure to search through the possible conformations in order to find the global minimum (or other acceptable solution). Finally, the conformational space—that is, the range of conformations that can be constructed and the means to transform one conformation into another—must be specified in order to constrain the search. Clearly, these elements are not independent and must fit together in order to form a coherent model. For example, an energy function need not evaluate a conformation that is not part of the allowable space. Nonetheless, each of the three components offers a different means to incorporate empirical information into the simulation.

B. Geometrical Representations

In order to reduce the number of degrees of freedom, most simulations use a reduced model description of the protein in which only a subset of the atoms are present. There are many variations on this theme, most of which have been previously reviewed [26]. The most common approach is to represent the main-chain N–C_α–C' atoms explicitly, with the side chain either being represented by the C_β atom or by an extended model atom corresponding to the approximate center of mass of the side chain. The bond distances and bond angles are usually fixed to standard values, thereby leaving the backbone dihedral angles ϕ and ψ as the only degrees of freedom (with the conjugated peptide dihedral angle fixed at 180°, in some cases allowing 0° as well for proline residues). The dihedral angles can either be restricted to a limited number of allowable conformations or be allowed to continuously vary within a specified region, and both of these approaches have been explored in our group and others.

Another method we are currently developing divides the molecule into segments based on the assigned secondary structure. The relative positions of the segments and the positions of the residues within each segment are optimized in distinct steps, thereby allowing the overall topology to evolve using a long-range potential with the detailed atomic coordinates to be adapted accordingly. The protein backbone is initially not required to be continuous from one segment to the next; and each segment can be deformed as the topology changes, creating unnatural bond lengths and angles. The correct covalent connectivity, rather than being rigid from the start, is gradually annealed in using a special constraint potential during the course of the simulation.

The details of side-chain conformation are generally determined by local interactions and have relatively little influence on the overall topology of the fold. Methods have been developed to assign probable side-chain conformations based on backbone dihedral angles and observed preferences in the PDB, and this technique has been shown quite effective in correctly placing side-chain atoms on a fixed backbone [27]. The task becomes more difficult if there are

significant deviations in the backbone, because the details of the side-chain contacts will no longer be the same [28]. In a recent approach, the side-chain conformations are represented by specifying a distribution of discrete rotamer states without actually including any additional coordinates. The ability of the backbone conformation to adequately accomodate the side chains can be evaluated using a rotamer-dependent mean-field energy and a conformational entropy [29].

C. Search Algorithms

The most common minimization technique is based on the principle of simulated annealing, which involves generating an ensemble of structures which is slowly converged toward the lowest-energy region of the conformational space. This method requires that the conformational sampling be able to avoid becoming trapped in a local minimum, and a number of techniques have been developed to overcome this problem [9,30]. Other successful approaches include using a branch-and-bound algorithm to limit the scope of local searches [31], as well as combining discrete Monte Carlo trial moves with local gradient minimizations [32].

Lattice models have also been used in order to discretize the conformational space in three dimensions. A relatively fine-grained model can be searched using methods similar to those described above [33], or a coarser model can be used to generate a set of possible topologies which can then be further refined using a more detailed model [34]. Further refinement can be carried out by using consensus inter-residue contacts from simulations to generate new structures that attempt to reproduce as many as possible [35,36]. Searches can even be carried out directly in terms of inter-residue contacts and then used to generate three-dimensional coordinates [37]. Another means to simplify the conformational search is to increase the range of the potential interactions during the simulation in order to build up larger-scale features of the structure [38].

Our approach is the hierarchical algorithm [39,40], in which trial moves are generated and evaluated in three different steps. At the simplest level, segments of three residues (triplets) are generated by choosing three sets of (ϕ, ψ) values at random from an allowed list. Each triplet is immediately accepted or rejected according to whether or not the orientation of its endpoints falls into an allowed region of triplet conformational space. The second level consists of complete loop segments as determined by the secondary structure. These loops are evolved from previously existing structures by using the set of triplets from the first level as trial moves and by evaluating new loops based on the difference in overall geometry from the starting loop. The final level then corresponds to the entire molecule, for which the trial moves consist of substituting entire loops with the new loops generated in the second level. It is only at this final level that the structure is evaluated by calculating the full scoring function, which is then

minimized using a genetic algorithm consisting of separate mutation, hybridization, and selection steps.

D. Scoring Functions

In most current prediction methods, the objective of the scoring function is not to reproduce the physical properties of the system, but to provide the best possible recognition of the native structure. These functions can be parameterized strictly on a statistical basis to optimize their performance [41]. Although there is some correlation between statistical potentials and those developed from physical principles [42], the former generally provide better results for predictions [43]. The energetic point of view is often used to motivate the development of a scoring function, but in practice the goal is simply to evaluate the relative probability that a given structure corresponds to a real protein. A typical energy can be defined as

$$E = \sum_{ij} E_{ij}$$

where the pairwise residue–residue energy is

$$E_{ij} = -kT_0 \ln P_{ij}(r_{ij})$$

and P_{ij} is the relative probability of finding residue pair i–j at a distance r_{ij}. If one then uses the Metropolis test to accept or reject a trial move from initial energy E_i to final energy E_f according to the value of $\exp(-(E_f - E_i)/kT)$, the same algorithm could be equivalently formulated in terms of accepting moves with a probability of $(P_f/P_i)^{\alpha}$, where $\alpha = T/T_0$ and

$$P = \prod_{ij} P_{ij}(r_{ij})$$

In principle, one could try to maximize the probability, its logarithm, or for that matter any other monotonic function of it.

Empirical scoring functions generally consist of multiple components, both sequence-independent and sequence-dependent [44,45]. The former include terms to control the overall size and shape of the molecule, as well as characteristic features of local structure depending on the geometrical model being used, whereas the latter take into account the specific interactions among residues. Some scoring functions are based on physical principles, such as electrostatic interactions [38] and van der Waals forces [46], with additional parameterization based on the PDB. The most common type of scoring function, however, is based directly on observed distances between different amino acid

pairs in the PDB, and it is formulated as a table of (possibly distance-dependent) pairwise contact probabilities between amino acid types [47,48]. They differ mainly in the functional form to which they are fit, as well as in the details of the normalization of the probabilities, which is a nontrivial task for a heterogeneous data set like the PDB [48,49]. The scoring functions used in our group are of this type, the details of which have been published elsewhere [32,39].

Additional specificity can be built into the scoring function in several ways. Specialized pattern-recognition and multibody terms can be included to generate more realistic secondary and supersecondary structural motifs [45,50]. The secondary structure can also be explicitly taken into account when calculating residue contact probabilities, in order to distinguish interactions between amino acids in different secondary-structure units [51]. In a more sophisticated approach, the local sequence homology is used to adjust the statistics for a particular target sequence [19]. The trend toward more explicit pattern recognition and sequence specificity in the generation of scoring functions allows more of the subtle homologies in the PDB to be exploited, although some chemical insight is still required to express it in an appropriate functional form.

III. CONSTRAINT METHODS

Constraints provide a very direct means to add information to a simulation—simply requiring all generated structures to satisfy certain additional conditions. This approach has been used extensively to generate three-dimensional structures from NMR spectra [52], which provide data in the form of interatomic distances. In principle, if one had enough distance constraints, the problem would be overdetermined and could be solved mathematically with no further information required. It has been shown, however, that the use of knowledge-based simulations based on homologous structures or fragment libraries from the PDB provides more accurate models than constraint-based methods alone [20,53].

In the case where the constraints alone are insufficient to determine the structure, they can still be used to supplement energy-based simulations. The goal in this case is to make the most effective use of the constraint information and to obtain good results with a minimum of additional information required. Because the source of the constraints is typically experimental spectra that must be assigned and interpreted, or theoretical methods (such as multiple sequence alignments) that may be incorrect, it is also important to take into account errors especially in difficult cases where the input data is incomplete or uncertain. Under these conditions, the constraints can be regarded as an additional component of the scoring function, expressing the probabilities of different structures, rather than as a rigid requirement. In many implementations, these interpretations are in fact equivalent.

A. Types of Constraints

Constraints can in principle be applied to any property of the structure where some sort of prefered value can be determined; however, the most common are those that correspond to experimental information. Some common types are outlined in the following sections.

1. Distance Constraints

Although the use of distance constraints to determine structures from NMR spectroscopy is well-established [52], these are experimentally determined structures rather than predictions in the sense used here. Applying a limited number of distance constraints to the simulation of an unknown structure in order to determine the gross topology rather than the detailed coordinates is a more recent approach [54]. This work showed, however, that the number of distances required for this purpose was at least an order of magnitude less than that needed for a complete structure determination. The emphasis in recent years has therefore been to reduce this number even further and to increase the size of protein that can be studied, with the goal of obtaining better structural information while requiring fewer experiments. In practice, tests are usually carried out on known structures where a given number of distances can be chosen at random to simulate such data.

2. Angle Constraints

There are currently experimental techniques to extract dihedral angles from NMR chemical shifts and coupling constants [55,56]. There is, however, a considerable margin of error on the order of $\pm 45^\circ$ in the actual values, which varies according to secondary structure [57]. These values are therefore insufficient for purposes of constructing the backbone by a sequential buildup; however, the target values and corresponding uncertainties can be applied as constraints in a torsional scoring function. The same applies to local backbone distance constraints, which in a reduced model are more conveniently expressed as limits on the dihedral angles rather than as specific interatomic distances. Although the dihedral angles in principle determine the structure directly, it is possible to have significant local variations in ϕ and ψ without appreciably changing the overall fold. The goal is therefore to use local dihedral constraints to bias the simulation toward the native structure while maintaining sufficient flexibility to avoid propagating errors due to incorrect values. Angle constraints can also be effectively combined with distance constraints to obtain greater precision from experimental data [58].

3. Other Types of Constraint

Data from NMR experiments which measure residual dipolar coupling [59] and paramagnetic relaxation [60] can be used to derive long-range geometrical

constraints and global features of the structure. These methods allow one to determine the relative orientation of N–H bonds relative to a common (unknown) reference frame, although not directly to one another. Although it is difficult to extract detailed information from this type of data due to the inherent degeneracy of the relative orientations, it is complementary to the types of constraints mentioned above and therefore can be very useful in folding simulations to screen out incorrect structures. This type of constraint lends itself well to a scoring-function approach, because it is easier to calculate the values that would be produced by a predicted structure and compare them with the experimental data than to impose *a priori* constraints in generating the structure. Although this type of constraint shows considerable promise, its use in simulating larger proteins is still less well developed than the more traditional distance and angle constraints.

B. Deriving Constraints from Predictions

Although the emphasis so far has been mostly on experimentally determined constraints, the same techniques that have been developed, especially in the case of uncertain or ambiguous constraints, can be just as well applied to theoretically predicted data. In cases where this is derived from sequence homology and/or multiple sequence alignments, the use of predicted constraints effectively generates a sequence-specific scoring function where any additional information is added to the generic scoring function already in place. Probable contacts can be derived from correlated mutations in a family of aligned sequences [18,61]. If a structure is known for at least one member of the family, contacts that are observed in the known structure which are likely to be conserved can be identified by looking at correlated mutations across the sequences, using the hypothesis that pairs of sites which have an increased probability of changing in concert are more likely to be in physical contact. Because there is a large number of possible pairs in a given sequence, as well as a relatively low signal-to-noise ratio in evaluating correlations, this method is less effective when based solely on sequence data without a reference to identify pairs that are likely to be in contact at all. On the other hand, extracting probable contact pairs can provide better results than direct homology modeling when the homology is weak and the structural alignment is uncertain.

Probable backbone dihedral angles can be predicted using sequence-based methods similar to those used in predicting secondary structure [62,63]. Although this could be considered a simple torsional potential, it is included in this section because it nonetheless incorporates sequence-specificity into the potential and can be implemented using the techniques of flexible angle constraints. In another method, contact distances between residues in different helices were determined by first selecting likely hydrophobic residues to form helix–helix contacts and then using a distance range typical of observed helix

pairs in the PDB [36]. Distance constraints can also be generated directly from the simulation results themselves [35]. In an ensemble of predicted structures, the frequencies of inter-residue contacts can be analyzed to identify those that are observed across a range of structures. These ''consensus'' contacts can then be imposed as constraints and used to generate structures that are better than any of those used to derive the constraints. A similar approach has been used in our group to correctly identify inter-residue contacts using an ensemble of structures in which no structure individually had the correct topology.

C. Constraint Implementation

Constraints are typically applied as a penalty function that is added as an extra term in the scoring function, often as some simple function (e.g., harmonic) of the difference between the actual and target values. Other strategies are possible, however, and constraints have also been used systematically in the construction of model structures. This can be applied to distance constraints, where a buildup procedure is used to generate structures that satisfy all constraints [64]. Angle constraints can also be used to systematically search the conformational space, both using a branch-and-bound procedure [65] or in a tree-search algorithm in combination with distance constraints [66].

In the case of sparse constraints, however, it has been shown that there is an advantage to using more flexible, or ''floppy'' constraints that allow for a more effective conformational search [67]. In our work, we apply inter-residue constraints to the C_β–C_β distances, regardless of the atoms involved in the original data. This is partly due to the practical problem of not representing side-chain atoms, but it also serves to simplify the calculation. The range of possible C_β–C_β distances consistent with the data is accounted for by using generous limits on the constraints. Rather than corresponding to a loss of precision, this actually improves the efficiency of the minimization.

We have studied a variety of functional forms for the constraint penalty functions and have found that a flat-bottom well with an exponential tail provides the best results. This penalty function has the form

$$U(r) = \begin{cases} -1, & r < c \\ -\exp(-r/d), & r > c \end{cases}$$

where c is the maximum constraint distance and d is the width of the tail. The best results are obtained with a square-well width of 8 Å and a tail width of 3 Å. The width is held constant independent of the actual constraint distance, because this allows greater flexibility and gives better scores to nearly correct structures. In fact, even for distances known to be less than 6 Å, setting c to 8 Å gave better results than a c of 6 Å, due to the fact that correct contacts are better recognized

despite local errors in the structure. For the same reason, no inner cutoff was used other than the usual excluded volume term.

In cases where the constraints are known to be accurate, good results can also be obtained for penalty functions that become large at long distances, such as linear or quadratic tails. This gives a large energy for any structure that severely violates any constraint. This is fatal, however, in cases where some constraints are incorrect or even contradictory. It is therefore important to ensure that while there is a favorable score for satisfied constraints and an attractive force in their vicinity, in the limit of grossly violated constraints the corresponding score goes to zero and is simply ignored.

1. Ambiguous Constraints

Ambiguous constraints arise in working with NMR NOE data that haven't been completely assigned [68]. In cases where similar residues have virtually the same chemical shifts, it can be difficult to identify which sites in the sequence are responsible for an observed contact. The same principle also applies to cysteine (S–S) linkages where several different pairings of cysteine residues may be possible. In such cases, carrying out a simulation with simultaneous constraints corresponding to each possibility can be used to determine the correct pairings [69]. The results of simulations with conflicting distance constraints have even been used to eliminate incorrect assignments for subsequent simulations and eventually deduce the correct contacts [21,70]. Another approach that gives rise to ambiguous constraints is the simulation of predicted secondary structure, where the different possible assignments can be expressed as a weighted combination of short-range distance constraints [71].

In our implementation, ambiguous distance constraints are simply expressed as a linear combination of all possibilities; in other words, all constraints are treated equally. As the penalty function goes to zero for violated constraints, the score is essentially the same for a residue that satisfies any one of the possible constraints, and the structure as a whole is optimized to satisfy as many as possible. An optional weighting factor can be included to represent the relative probabilities associated with different assignments.

D. Results

In order to test some of the ideas discussed above, we have carried out a number of experiments on known structures by artifically generating constraints from the PDB coordinates. Although this is far removed from real-world applications, having precise control over the quantity and quality of the supplemental data allows the methods to be carefully evaluated and allows their limits to be better determined. In the following sections, some representative examples are presented to illustrate the progress that has been made, and comparisons are shown with similar work from other groups.

TABLE I
Results of Simulations with Constraints for 3ICB

	Constraints	Low RMS	Average RMS	Standard Deviation
Present work	0	4.6	9.8	1.9
	10	3.0	4.9	1.3
	89	3.0	3.3	0.2
Aszódi et al.	0		10.0	1.5
	10		6.3	2.0
	86		2.9	0.2

1. Distance Constraints

The implementation of distance constraints was tested using two small globular proteins that have been previously studied in the literature: calcium-binding protein (3ICB), an α protein with 72 residues, and tendamistat (3AIT), a β protein with 62 residues [72]. In each case, a total of 10 constraints were chosen at random from among the eligible pairs of residues in the crystal structure. This was repeated for 20 simulations, each using a different set of constraints, and compared with earlier literature results [73]. The results are summarized in Tables I and II. For 3ICB, 10 constraints are sufficient to find as good a structure as was found using all of the constraints. Because of the use of ideal β-strands without any sort of strand-pairing potential, 3AIT proved to be much more difficult, although the addition of 10 constraints does also lead to a significant improvement. Other published simulations [74] show better results when all of the constraints are used, but fail completely for small numbers of constraints. A test was also carried out with a larger molecule, myoglobin (1MBA), an α protein with 140 residues, the results of which are shown in Table III. Using 20 constraints in this case, a structure with an RMS deviation of 4.5 Å was obtained, comparable to 4.9 Å reported elsewhere for the same set

TABLE II
Results of Simulations with Constraints for 3AIT

	Constraints	Low RMS	Average RMS	Standard Deviation
Present work	0	8.4	9.7	0.4
	10	4.8	8.4	1.3
	116	3.6	6.8	1.6
Aszódi et al.	0		9.4	0.7
	10		5.8	0.6
	120		3.7	0.2

TABLE III
Results with Constraints for 1MBO

	Constraints	Low RMS	Average RMS	Standard Deviation
Present work	0	7.1	12.3	1.8
	20	4.5	10.3	1.8
	30	3.2	5.7	1.2
	50	3.6	5.3	1.6
	100	2.9	4.5	1.0
Skolnick et al.	20	4.9	5.6	

of constraints [75]. This result improved to 3.2 Å with a random selection of 30 constraints, which was essentially equivalent to results obtained with larger numbers of constraints.

2. *Angle Constraints*

Within our hierarchical model, it is more convenient to implement angle constraints in a different manner. Instead of using a scoring-function approach, we introduce the constraint information at the level of the list of allowed ϕ–ψ pairs. Because the pairs are selected randomly, the number of values in each region will determine the corresponding bias in the simulation. Test calculations were carried out for myoglobin (1MBO) in which part of the dihedral list corresponded to the usual distribution and the other part was limited to a region with a width of 30° around the target values. Clearly, if the weight of the latter region is 100%, this represents a rigid constraint, however, in order to maintain the flexibility of the simulation and allow for the possibility of incorrect data, it is useful to retain some of the original distribution. Simulation results are summarized in Table IV as a function of the relative weight of the constraint region. Good results are obtained with a 50% weighting, indicating that there is

TABLE IV
Results of Simulations of 1MBO Using Angle Constraints with Different Relative Weights

Constraint Weight (%)	Low RMS	Average RMS	Average Score
0	8.1	11.1	−172
6	7.4	11.7	−172
20	4.9	9.8	−173
30	5.1	6.5	−218
50	2.5	4.1	−226
100	1.7	2.7	−226

TABLE V
Results for 1MBO with 100 Total Constraints

Number of Good Constraints	Number of False Constraints	Low RMS	Average RMS
100	0	2.6	4.7
75	25	3.7	5.2
50	50	4.0	6.8
30	70	5.3	8.9
20	80	6.0	10.8

a strong cooperative selection. On the other hand, a control experiment was carried out also with 50% weighting, in which the target values were chosen at random, thus giving a nonsensical structure if taken together. The results in this case were essentially the same as those with no constraints at all, showing that the simulation is nonetheless able to ignore incorrect data.

3. Ambiguous Constraints

In order to test the sensitivity of the simulation with respect to incorrect data, a series of experiments was carried out in which the total number of distance constraints was held fixed, but the number of which were correct was varied. In a first trial, again with myoglobin (1MBO), 100 constraints were used. The correct constraints were derived by randomly selecting from among the possible contacts observed in the PDB structure, and the remaining number were randomly selected from pairs of residues known to be at least 20 Å apart in the correct structure. This was repeated with several different sets of constraints, to avoid any bias due to a lucky choice of correct constraints. The results are shown in Table V. Compared with the results in Table III, there is clearly a loss in performance due to the presence of incorrect constraints; however, reasonable results can still be obtained in cases where the nonsensical constraints actually outnumber the real ones. A similar experiment using flavodoxin (2FX2), a mixed α/β protein with 143 residues, is shown in Table VI. Although there is an increasing number of misfolded structures, as indicated by the average RMS

TABLE VI
Results for 2FX2 with 100 Total Constraints

Number of Good Constraints	Number of False Constraints	Low RMS	Average RMS
100	0	4.6	7.2
75	25	5.2	9.4
50	50	5.2	11.9

TABLE VII
Results for 1MBO with 150 Total Constraints

Number of Good Constraints	Number of False Constraints	Low RMS	Average RMS
100	50	4.0	5.9
50	100	3.7	7.7
30	120	6.1	10.7
20	130	(9.2)	(13.2)

deviation, the simulation is still able to find reasonable structures with only half of the constraints correct. A further experiment on myoglobin with 150 constraints, shown in Table VII, shows that the constraints remain useful with as few as 20% correct. Values in parentheses are actually higher than in a comparable simulation with no constraints at all. These results support the idea that, up to a certain limit, more data is better even if it becomes less reliable.

4. Predicting Constraints

The most promising method for predicting distance constraints is based on correlated mutations in multiply aligned sequences. This approach has been used in folding simulations with on average about 25% of tertiary contacts predicted to within ± 1 residue in the sequence, and it was shown that this is sufficient to generate reasonable fold predictions [18,61]. In experiments carried out in our group, summarized in Table VIII, the predicted constraints were found to be more than sufficient to generate reasonable structures. Predictions in this case are considered correct if the two C_β atoms are in fact within the 8 Å

TABLE VIII
Contact Prediction Accuracy

Target:	1CCR	2LHB	1MIL
Sequence length:	107	134	84
Aligned sequences:	10	7	6
Maximum indentity:	62	31	29
	Low Sensitivity		
Predicted contacts:	88	81	84
Percent accurcy:	93	89	88
	High Sensitivity		
Predicted contacts:	33	47	45
Percent accuracy:	100	89	87

well used in the simulation. Results are shown for both low and high sensitivity, meaning that the criterion used to predict contacts based on the statistical significance of the sequence correlations was more strict in the latter case. Although this improves slightly the accuracy of the predictions, the larger number of total contacts provides a clear advantage for the low sensitivity predictions. In particular, in the case of 1MIL where there are relatively few aligned sequences with low homology, the selection criterion was of little use and yet the overall quality of the predictions was quite good.

IV. LIMITING THE SEARCH SPACE

Generic information about protein structure can be incorporated in a simulation by restricting *a priori* the conformations that can be generated. If the simulation is only capable of producing structures with certain realistic properties, the odds of finding the correct fold are greatly enhanced. In the extreme case, the choices would consist of a limited number of compact folded structures for the entire sequence. In such a ''simulation'' the global minimization problem is trivial (exhaustive enumeration becomes feasible) and the scoring function need only distinguish among topologicaly different structures without reproducing any of the interactions that stabilize such structures in the first place. Clearly, all the work is being done in the initial definition of possible trial structures, which therefore becomes the determining element of the algorithm. There is a necessary tradeoff between using the characteristics of known folds to limit the search and running the risk of incorrectly excluding a structure that had not been previously seen.

A trivial application of this principle, however, is the use in the hierarchical algorithm of a list of allowed ϕ–ψ pairs in generating new segments. This eliminates the need for a scoring function to penalize unfavorable regions of the Ramachandran map, as well as the need to sample such unlikely regions of the conformational space. Although the definition of this list is entirely empirical, based on observation of the PDB, it still represents real interactions that a new structure would be very unlikely to violate.

A. The Principle of Threading

The most obvious way to select realistic structures is to simply use those that are already known in the PDB, and this is the basis of what is commonly known as threading. Threading is normally associated with the problem of fold recognition—that is, identifying homologous structures in the PDB—rather than in the context of simulation. It is included here as the limiting case of a restricted search in order to establish a relationship between the *ab initio* and fold-recognition approaches and also to provide a framework for describing various intermediate methods that have been developed.

In its simplest incarnation, threading consists of attempting to map the target sequence onto the backbone coordinates of all structures in the PDB of equal or greater length. This can be visualized as stringing a flexible chain of amino acids along the fixed scaffold of a known structure—hence the name threading. In this simple approach, the number of possible alignments (mappings of the target sequence onto the corresponding residues of a known structure) is limited, and most empirical scoring functions are capable of recognizing truly homologous structures. The method fails, however, to identify distant structural homologs and is obviously incapable of generating any new folds. More realistic methods allow the connectivity of the template structures to be modified [76] and allow gaps and insertions to be introduced in the alignments. This, however, greatly increases the number of possible alignments and makes the problem of recognizing homologous structures that much more difficult [77].

B. Local Threading and Fragment Lists

One way to overcome the combinatorial problem is to divide the problem into smaller local alignments. This can be done as a first step in generating a global alignment to a single known structure [14], or alternatively to identify shorter segments that align to parts of different structures. The structure of a known fold can be described by specifying the local environment of each residue: secondary structure, polarity, and solvent exposure [78]. This allows the threading to be carried out locally, aligning a linear sequence to a series of profiles by the same methods used for sequence–sequence alignments, independently of the rest of the molecule.

The resulting local alignments lead to a large number of possible combinations that must still be reassembled into a single structure. In this situation, rather than attempting to either select the best local homologs or carry out an exhaustive enumeration, it is more effective to return to a stochastic simulation where the local templates act as lists of trial structures for each segment. In this way, the principle of using a restricted set of conformations can be extended across various levels of structure: from individual amino acids (as in a typical simulation) to multiresidue fragments, loop and secondary structure elements, supersecondary motifs, and ultimately entire domains (as in a typical threading calculation).

1. Using a Motif Library

A set of commonly occurring structural motifs, along with their associated sequence profiles (the probability for each amino acid to occupy each site in the structure), have been extracted from the PDB using local sequence and structure alignments [13]. Experimental evidence has even shown that some peptides do in fact adopt the corresponding motif structure in isolation and that strong fits to the sequence profile can possibly be used to identify sites of folding initiation

[79]. These motifs have been successfully used in a simulation algorithm to predict new folds, providing one of the more impressive achievements in the CASP-3 experiment (see Simons et al. in Ref. 8). This motif library has now been united into a global prediction scheme using a hidden Markov model to encode extended sequence profiles [63]. A similar library of loop motifs has also been used to model loop regions in homology models using the flanking secondary structure as a guide [80].

2. *Mapping Conformational Space*

Rather than attempt to identify common motifs, another approach is to try to identify a minimal set of building blocks that can be used to represent any known structure [12]. This essentially corresponds to a redefinition of the geometrical model in which the smallest unit of structure becomes a five- or six-residue fragment. The result of using this model is a greatly reduced number of degrees of freedom and a more efficient exploration of conformational space.

C. Fragment Screening and Enrichment

As an alternative to using preselected fragment lists to build up a model structure, a more general approach is to use the characteristics of homologous structures to screen possible conformations. The idea is still to allow arbitrary conformations, as in a traditional simulation, but to increase selectively the proportion of generated structures with the desired protein-like qualities. By using homologous motifs from the PDB to define the selection criteria, sequence-dependent conformational preferences can be introduced into the simulation without reducing the flexibility of the model.

1. *The Hierarchical Approach*

In the hierarchical algorithm [40], the structures of the residue triplets are generated from independent residue conformations which are determined by the three amino acid types. These triplets are then screened according to the relative orientations of the end residues, which determine the positions of the flanking segments. For a given target sequence, the distribution of triplet geometries is calculated for segments in the PDB which have a local sequence homology greater than a specified cutoff. This distribution is used to accept or reject randomly generated triplets so as to reproduce the observed probabilities of finding a triplet with a given geometry. This generates a sequence-specific list of triplet conformations which can then be used to generate larger fragments. In preliminary tests using this method on a set of test proteins, both the average energy and deviation from the native structures was found to decrease as the selectivity of the screening (the homology threshold) was increased. In these tests, any structure with significant global homology to the target sequence was excluded from the fragment database.

Loop segments of varying lengths are then built up by randomly selecting from the lists of triplet conformations. Loops are again selected by comparing the end-to-end distances and rotations with homologous loops in the PDB. Although the internal structure of the loops is free to vary, the goal is to generate structures that are more likely to produce a favorable positioning of the flanking segments. Because this type of selection is applied successively at three different levels of structure, the overall process is quite efficient and the cutoff parameters can be freely adjusted to give the desired level of structural similarity and sequence homology at various stages of the simulation.

D. Modeling Secondary Structure

Due to its well-characterized regular motifs, secondary structure is an obvious candidate for fragment-based modeling. Indeed, a common approach, and the one traditionally used in our group, is to simply hold the secondary structure fixed during the calculation, which is an extreme application of the principles described in this section. In cases where the secondary structure is predicted from the sequence, this is a crude application of fragment selection by sequence profile. This effectively removes a large number of degrees of freedom and eliminates the need to use the scoring function to stabilize α-helix and β-sheet conformations.

This approach can be generalized, and some flexibility reintroduced into the structure, by developing specific models to reproduce the observed variability within the regular structures. A list of strand or helix structures can be assembled from the PDB, with associated error tolerances on the dihedral angles to account for kinks and imperfections, and this can be used to define the possible conformations of an arbitrary helix as a single unit. Sheets are in general more complex and show more natural variability; however, the possible collective structures have been extensively studied and characterized [81,82]. Using the generic properties of β-sheets, a library of conformations with varying twist and curvature can be constructed for an arbitrary sequence.

As a preliminary test, we have carried out a series of simulations with a range of possible helix and strand geometries to determine if the tertiary contacts would be sufficient to identify the native structure. The list of trial structures consisted of a continuous deformation from an ideal geometry to the (known) native geometry, with the same deformation vector extended to also generate even more deformed structures. The results for a set of test sequences are shown in Figs. 1 and 2 for helices and sheets, respectively. The deformations are grouped into discrete bins, and in each case the corresponding native structure falls into bin number six. For helices, which have a smaller average deformation, the distribution is relatively smooth with a maximum at the native geometry. In the case of β-strands, the distribution is more-or-less flat with a

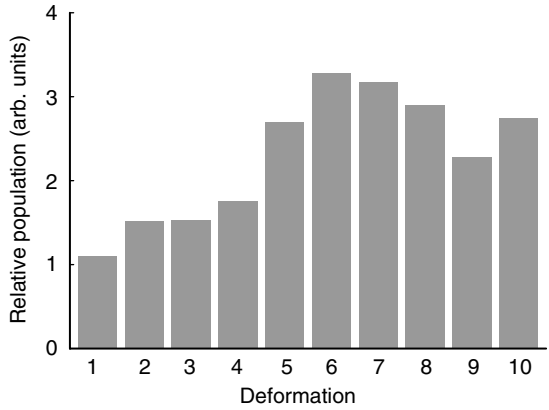

Figure 1. Helix selection frequency as a function of relative deformation for a set of test proteins. In each case, bin 1 corresponds to an ideal structure, bin 6 corresponds to the native structure, and the other bins correspond to a linear extrapolation.

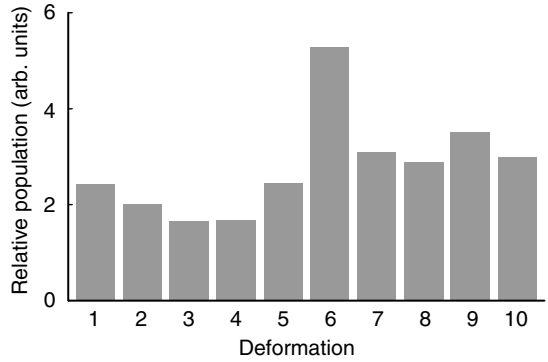

Figure 2. Strand selection frequency as a function of relative deformation for a set of test proteins. The bin deformations are as in Fig. 1.

more pronounced peak at the native geometry, suggesting that the correct deformation does in some sense better "fit together" and is energetically favored.

V. HOMOLOGY AND STRUCTURAL TEMPLATES

For homologous proteins, a threading alignment as described in the previous section can be used to provide a template for the entire structure. In the absence of global homology, however, local alignments can still be used to extract

localized structural constraints. This approach, unfortunately, results in the loss of any information about the overall topology of the tertiary structure, which is the most difficult part of the folding problem. An alternative is try to identify a smaller number of aligned residues, possibly with significant gaps, in order to provide key reference points for determining the overall structure. In this case, much local information will be missing, and local structure must still be determined using standard simulation methods; however, the relative three-dimensional positions of different parts of the structure can be controlled. This is consistent with the chemical interpretation of a relatively small number of conserved residues playing an important role in both fold stability and function (although of course there are many exceptions to this picture.) When the homology is weak, it may be more effective to try to identify the most probable conserved residues than to rely on a global alignment that is likely incorrect.

A. Identifying Structural Templates

The most straightforward approach is to carry out a standard threading calculation and exclude regions with a poor alignment score. Template residues can also be excluded in regions where the target is not predicted to have a regular secondary structure, or where the template secondary structure differs from that predicted for the target. In this way, the parts of the alignment most likely to correspond to a stable core can be identified and the simulation can be used to fill in the gaps. In our implementation, the superposition of the selected residues with their corresponding coordinates in the aligned template structure is then used as an additional contribution to the scoring function. Another approach is to constrain the simulation to follow the template structure, but to allow the specific alignment to change during the simulation [83].

Positions likely to be conserved in a sequence can also be identified by searching through a database of known sequence patterns such as PROSITE [84]. In our approach, patterns identified in the target sequence were then used to search the PDB for structures containing the same patterns. The coordinates of the conserved residues were then averaged over all matching structures to generate a composite template that was then used in the simulation. An experiment was carried out for the myoglobin sequence (1MBO) using coordinates from seven structures in the PDB having less than 20% sequence identity with 1MBO to obtain the template coordinates. The results of the simulation are shown in Table IX as a function of the number of template sites used. Good structures were obtained using a template with a relatively small number of aligned residues.

B. Multiple Templates

In many cases there may be more than one possible template for a given target sequence. This can arise from different choices of reference structure, or for the

TABLE IX
Performance as a Function of Template Size for 1MBO

Size of Template	Low RMS	Average RMS
0	9.5	13.8
11	4.7	9.4
25	3.0	4.2
50	2.4	3.3
146	2.1	3.0

same reference structure different alignments and choices of predicted secondary structure. In addition, different parts of the target sequence might align well to different structures in the PDB. In such cases, the simulation can be used to choose among conflicting alignments and to combine different templates.

1. *Template Competition*

In our implementation, the same philosophy is used as in the case of distance constraints. The scoring function is a spline-fit switching function of the actual superposition RMS deviation with the template coordinates. This function is equal to -1 below a lower cutoff value, equal to zero above an upper cutoff value, and varies smoothly in between the two. Conflicting templates can therefore be used simultaneously, and a favorable score will be obtained for structures that superpose well on any one or more of them and no penalty is assessed for distant templates. The simulation can therefore be used to identify which of the possible templates gives the best fit consistent with the connectivity of the sequence and the generic scoring function.

2. *Results*

This methodology was used in the most recent CASP experiment, from which two representative examples will be described which illustrate how the methodology was applied. For sequence T0089, threading results suggested eight possible templates for the N-terminal region, four possible templates for the C-terminal region, and three or four different alignments and secondary-structure assignments in each case. None of the alignments had a sequence identity greater than 15%, and in addition there was a gap of about 120 residues between the two templates. Simulations were run using all possible combinations of two templates, and the final prediction was selected based on the fit to the templates, the overall energy, and the ability of the connecting segment to fold.

The situation was reversed in the case of sequence T0087, where instead of a gap there was an overlap of over 100 residues between the two proposed templates. In this case, 11 choices for the N-terminal region and six choices for

the C-terminal region were identified, all with about 10–15% sequence identity. For each possible combination the two templates were used simultaneously, thereby generating a conflicting set of constraints for the region in which the two overlapped. The final prediction was selected as that which provided the best simultaneous fit for both templates, thus hopefully giving a relative orientation of the two domains consistent with the context of each.

Unfortunately, the preliminary results indicate that none of the proposed templates was correctly aligned to the native structure, so it is difficult to judge the performance of the simulation methodology. In each case, however, the submitted structures were correctly ranked, with the best one selected as the first choice.

C. Local Templates

The use of multiple simultaneous templates can also be extended to model generic structural motifs. In contrast to the method of segment libraries discussed earlier, these are structural relationships which are nonlocal in sequence; rather than describing the local backbone conformation, the goal is to describe the relative spatial orientations of different structural elements. The use of multiple templates allows different possibilities to be considered, thereby providing a library of three-dimensional relationships. This use of generic structural templates provides a general alternative to local multibody scoring functions that recognize specific structural motifs.

1. β-Strand Pairing

Generating realistic β-sheet structures is a notoriously difficult problem due to the specific relative orientation of noncontiguous backbone segments produced by the H-bonding pattern. The H-bonds themselves, however, are short-range interactions that are difficult to simulate and often fail to produce the desired overall structure. Specific multibody interactions that take into account strand orientation are therefore often used to overcome this problem [45,85–87].

An alternative approach for correctly aligning two β-strands is to extract a template of a similar strand pair from the PDB, which can then be used to superimpose the target strands. A library of possible pairings can be generated based on sequence homology, and the technique of multiple templates described above can be used to select a suitable candidate for each interacting strand pair. To determine whether or not templates derived from unrelated structures could provide correct strand-pairing geometries, the closest structural homologs in the PDB were identified for a number of strand pairs, along with the best superposition in a list of the top 10 sequence homologs. Shown in Table X are the results of this experiment for the mixed α/β protein ribonuclease A (2RAT). (Sequences with more than 20% overall similarity to the target were excluded

TABLE X
Strand-Pairing Templates for 2RAT

Strand Pair	Length	Best Possible	Homologous
1–4	5, 8	0.64 Å (1BIA.1)	0.99 Å (1ZXQ)
4–5	8, 8	0.82 Å (1BYT)	1.93 Å (2MEV.2)
2–3	3, 3	0.08 Å (8FABA.A)	0.16 Å (2ENG)
3–6	3, 6	0.34 Å (1EFT)	0.81 Å (1BLI.A)
6–7	6, 8	0.91 Å (1A62)	2.56 Å (1CBJ.A)

from the calculation.) Reasonable models can be obtained for each pair, despite the lack of global homology.

2. *Hydrophobic Contacts*

A similar approach can also be applied to helix pairs, which, despite being linked only by hydrophobic contacts, tend to pack in well-defined relative orientations. It has been shown that by identifying conserved hydrophobic contacts between different helices, a model can be found in the PDB which reproduces the correct helix–helix packing and can be used to reconstruct the tertiary structure [88]. Because the helix structure is very regular, a single contact geometry is sufficient to generate a helix template of arbitrary length using a standard backbone conformation.

VI. NEW DIRECTIONS

The next logical step in the evolution of structure prediction is to generalize further the knowledge-based methods described so far in order to make maximum use of the motifs in the PDB, even in the absence of any detectable *a priori* homology, and to eventually replace the physically motivated idea of a universal energy function. Local structure will be modeled using fragment libraries, inter-residue interactions through generalized distance constraints, and multibody correlations through localized motif templates. The scoring function will become a moving target that adapts itself to the results of the simulation, adding a knowledge-based component to the already sophisticated search methods currently in use.

A. Sequence-Specific Potentials

Flexible distance constraints can be used to express the probability of forming different specific contacts in the structure, based on the context of each residue. Conceptually, if contact probabilities were to be predicted solely on the basis of amino acid type (hydrophobic residues are more likely to be in contact with other hydrophobic residues), this simply reduces to a traditional generic energy

function. Pair potentials have already been developed which derive contact probabilities based on local sequence [19], local secondary structure [51], and β-strand pairing [50]; any other observed correlation can be combined and expressed in the same way. Generalized sequence-based methods (such as in Ref. 63) can also be used to derive sequence-specific scoring functions for local conformation and structural context, allowing for a customized selection of fragments and templates.

B. Constraint Refinement

The results of the simulation itself can also be used to improve the prediction of inter-residue contacts, thus allowing an iterative series of simulations to generate successively more specific scoring functions. This is analogous to the use of iterative simulations in assigning NOE signals in NMR spectroscopy [21], except with purely theoretical input. It has been shown, however, that the statistical analysis of an ensemble of predicted structures can be used to derive more accurate contact information than any of the structures individually [35]. Preliminary experiments in our group have shown that it is possible to start with a large number of possible contacts and, by successively eliminating those that are observed less frequently in the ensemble, to eventually identify the correct native contacts.

VII. CONCLUSION

Considerable progress has been made over the past few years in developing practical tools for structure prediction. Geometrical models, empirical scoring functions, and global minimization algorithms have all evolved together to increase the efficiency and selectivity of simulation-based methods. Different techniques have advantages and disadvantages: Discretized models gain in sampling efficiency at the expense of resolution, template models carry more three-dimensional information, constraint-based methods are less sensitive to alignments, and so on. The result, however, is an increasingly complete spectrum of methods that are beginning to achieve meaningful results in a variety of real-world applications. As more and more information is being added to sequence and structure databases, there is every reason to expect this trend to continue.

Acknowledgments

The authors would like to thank Sabrina Bédard, Geneviève Dufresne, Martin Éthier, Olivier LeBel, Éric Morneau, and Marc-André Thivièrge, who contributed to some of the results discussed here.

References

1. A. R. Fersht, *Curr. Opin. Struct. Biol.* **7**, 3 (1997).
2. H. S. Chan and K. A. Dill, *Proteins* **30**, 2 (1998).

3. N. D. Socci, J. N. Onuchic, and P. G. Wolynes, *Proteins* **32**, 136 (1998).
4. B. Honig, *J. Mol. Biol.* **293**, 283 (1999).
5. D. J. Bicout and A. Szabo, *Protein Sci.* **9**, 452 (2000).
6. J. Moult, *Curr. Opin. Biotechnol.* **10**, 583 (1999).
7. D. J. Osguthorpe, *Curr. Opin. Struct. Biol.* **10**, 146 (2000).
8. Third Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, *Proteins Suppl.* **3** (1999).
9. J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy, J. A. Saunders, K. D. Gibson, and H. A. Scheraga, *Int. J. Quant. Chem.* **77**, 90 (2000).
10. L. Jaroszewski, L. Rychlewski, B. Zhang, and A. Godzik, *Protein Sci.* **7**, 1431 (1998).
11. M. J. Schoonman, R. M. Knegtel, and P. D. Grootenhuis, *Comput. Chem.* **22**, 369 (1998).
12. C. Micheletti, F. Seno, and A. Maritan, *Proteins* **40**, 662 (2000).
13. C. Bystroff and D. Baker, *J. Mol. Biol.* **281**, 565 (1998).
14. R. Thiele, R. Zimmer, and T. Lengauer, *J. Mol. Biol.* **290**, 757 (1999).
15. Y. Wang, H. Zhang, W. Li, and R. A. Scott, *Proc. Natl. Acad. Sci. USA* **92**, 709 (1995).
16. M. Vendruscolo and E. Domany, *J. Chem. Phys.* **109**, 11101 (1998).
17. O. Olmea, B. Rost, and A. Valencia, *J. Mol. Biol.* **293**, 1221 (1999).
18. A. R. Ortiz, A. Kolinski, and J. Skolnick, *J. Mol. Biol.* **277**, 419 (1998).
19. J. Skolnick, A. Kolinski, and A. Ortiz, *Proteins* **38**, 3 (2000).
20. B. L. Podlogar, G. C. Leo, P. A. McDonnell, D. A. Loughney, G. W. Caldwell, and J. F. Barrett, *J. Med. Chem.* **40**, 3453 (1997).
21. M. Nilges, M. J. Macias, S. I. O'Donoghue, and H. Oschkinat, *J. Mol. Biol.* **269**, 408 (1997).
22. D. T. Jones, *Curr. Opin. Struct. Biol.* **10**, 371 (2000).
23. C. L. Brooks III, *Curr. Opin. Struct. Biol.* **8**, 222 (1998).
24. X. Daura, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark, *J. Mol. Biol.* **280**, 925 (1998).
25. Y. Duan and P. A. Kollman, *Science* **282**, 740 (1998).
26. R. A. Friesner and J. R. Gunn, *Annu. Rev. Biophys. Biomol. Struct.* **25**, 315 (1996).
27. R. L. Dunbrack, Jr., and F. E. Cohen, *Protein Sci.* **6**, 1661 (1997).
28. E. S. Huang, P. Koehl, M. Levitt, R. V. Pappu, and J. W. Ponder, *Proteins* **33**, 204 (1998).
29. A. S. Lemak and J. R. Gunn, *J. Phys. Chem. B* **104**, 1097 (2000).
30. H. A. Scheraga and M. H. Hao, *Adv. Chem. Phys.* **105**, 243 (1999).
31. D. M. Standley, V. A. Eyrich, A. K. Felts, R. A. Friesner, and A. E. McDermott, *J. Mol. Biol.* **285**, 1691 (1999).
32. V. A. Eyrich, D. M. Standley, and R. A. Friesner, *J. Mol. Biol.* **288**, 725 (1999).
33. A. Kolinski and J. Skolnick, *Proteins* **32**, 475 (1998).
34. Y. Xia, E. S. Huang, M. Levitt, and R. Samudrala, *J. Mol. Biol.* **300**, 171 (2000).
35. E. S. Huang, R. Samudrala, and J. W. Ponder, *Protein Sci.* **7**, 1998 (1998).
36. E. S. Huang, R. Samudrala, and J. W. Ponder, *J. Mol. Biol.* **290**, 267 (1999).
37. M. Vendruscolo, E. Kussell, and E. Domany, *Fold. Des.* **2**, 295 (1997).
38. F. Avbelj and L. Fele, *Proteins* **31**, 74 (1998).
39. J. R. Gunn, *J. Phys. Chem.* **100**, 3264 (1996).
40. J. R. Gunn, *J. Chem. Phys.* **106**, 4270 (1997).

41. D. J. Ayers, T. Huber, and A. E. Torda, *Proteins* **36**, 454 (1999).
42. D. Mohanty, B. N. Dominy, A. Kolinski, C. L. Brooks III, and J. Skolnick, *Proteins* **35**, 447 (1999).
43. T. Lazaridis and M. Karplus, *Curr. Opin. Struct. Biol.* **10**, 139 (2000).
44. K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, and D. Baker, *Proteins* **34**, 82 (1999).
45. A. Kolinski, B. Ilkowski, and J. Skolnick, *Biophys. J.* **77**, 2942 (1999).
46. A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Oldziej, R. J. Wawak, S. Rackovsky, M. R. Pincus, H. A. Scheraga, *J. Comput. Chem.* **19**, 259 (1998).
47. M. J. Sippl, *Curr. Opin. Struct. Biol.* **5**, 229 (1995).
48. R. L. Jernigan and I. Bahar, *Curr. Opin. Struct. Biol.* **6**, 195 (1996).
49. F. Seno, A. Maritan, and J. R. Banavar, *Proteins* **30**, 244 (1998).
50. H. Zhu and W. Braun, *Protein Sci.* **8**, 326 (1999).
51. C. Zhang and S.-H. Kim, *Proc. Natl. Acad. Sci. USA* **97**, 2550 (2000).
52. A. T. Brünger, G. M. Clore, A. M. Gronenborn, and M. Karplus, *Proc. Natl. Acad. Sci. USA* **83**, 3801 (1986).
53. L. Kirnarsky, O. Shats, and S. Sherman, *J. Mol. Struct. (Theochem.)* **419**, 213 (1997).
54. M. J. Smith-Brown, D. Kominos, and R. M. Levy, *Protein Eng.* **6**, 605 (1993).
55. M. Hong, J. D. Gross, W. Hu, and R. G. Griffin, *J. Magn. Res.* **135**, 169 (1998).
56. K. Kloiber and R. Konrat, *J. Biomol. NMR* **17**, 265 (2000).
57. R. D. Beger and P. H. Bolton, *J. Biomol. NMR* **10**, 129 (1997).
58. M. J. Bayley, G. Jones, P. Willett, and M. P. Williamson, *Protein Sci.* **7**, 491 (1998).
59. G. M. Clore, A. M. Gronenborn, and N. Tjandra, *J. Magn. Res.* **131**, 159 (1998).
60. J.-C. Hus, D. Marion, and M. Blackledge, *J. Mol. Biol.* **298**, 927 (2000).
61. A. R. Ortiz, A. Kolinski, and J. Skolnick, *Proteins* **30**, 287 (1998).
62. H. S. Kang, N. A. Kurochkina, and B. Lee, *J. Mol. Biol.* **229**, 448 (1993).
63. C. Bystroff, V. Thorsson, and D. Baker, *J. Mol. Biol.* **301**, 173 (2000).
64. D. A. Debe, M. J. Carlson, J. Sadanobu, S. I. Chan, and W. A. Goddard III, *J. Phys. Chem. B* **103**, 3001 (1999).
65. J. L. Klepsis, C. A. Floudas, D. Morikis, and J. D. Lambris, *J. Comput. Chem.* **20**, 1354 (1999).
66. G. P. Gippert, P. E. Wright, and D. A. Case, *J. Biomol. NMR* **11**, 241 (1998).
67. D. Bassolino-Klimas, R. Tejero, S. R. Krystek, W. J. Metzler, G. T. Montelione, and R. E. Bruccoleri, *Protein Sci.* **5**, 593 (1996).
68. M. Nilges, *J. Mol. Biol.* **245**, 645 (1995).
69. J. Boisbouvier, M. Blackledge, A. Sollier, and D. Marion, *J. Biomol. NMR* **16**, 197 (2000).
70. C. Mumenthaler, P. Güntert, W. Braun, and K. Wüthrich, *J. Biomol. NMR* **10**, 351 (1997).
71. C. C. Chen, J. P. Singh, and R. B. Altman, *Bioinformatics* **15**, 53 (1999).
72. J. R. Gunn, *Intell. Syst. Mol. Biol.* **6**, 78 (1998).
73. A. Aszódi, M. J. Gradwell, and W. R. Taylor, *J. Mol. Biol.* **251**, 308 (1995).
74. O. Lund, J. Hansen, S. Brunak, and J. Bohr, *Protein Sci.* **5**, 2217 (1996).
75. J. Skolnick, A. Kolinski, and A. R. Ortiz, *J. Mol. Biol.* **265**, 217 (1997).
76. F. A. Hamprecht, W. Scott, and W. F. van Gunsteren, *Proteins* **28**, 522 (1997).
77. X. De La Cruz and J. M. Thornton, *Protein Sci.* **8**, 750 (1999).

78. D. W. Rice and D. Eisenberg, *J. Mol. Biol.* **267**, 1026 (1997).
79. Q. Yi, C. Bystroff, P. Rajagopal, R. E. Klevit, and D. Baker, *J. Mol. Biol.* **283**, 293 (1998).
80. J. Wojcik, J.-P. Mornon, and J. Chomilier, *J. Mol. Biol.* **289**, 1469 (1999).
81. F. R. Salemme and D. W. Weatherford, *J. Mol. Biol.* **146**, 101 (1981).
82. F. R. Salemme and D. W. Weatherford, *J. Mol. Biol.* **146**, 119 (1981).
83. A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick, *Proteins* **37**, 592 (1999).
84. K. Hofmann, P. Bucher, L. Falquet, and A. Bairoch, *Nucleic Acids Res.* **27**, 215 (1999).
85. A. Monge, E. J. P. Lathrop, J. R. Gunn, P. S. Shenkin, and R. A. Friesner, *J. Mol. Biol.* **247**, 995 (1995).
86. T. Dandekar and P. Argos, *J. Mol. Biol.* **256**, 645 (1996).
87. D. M. Standley, J. R. Gunn, R. A. Friesner, and A. E. McDermott, *Proteins* **33**, 240 (1998).
88. M. Parisien, F. Major, and M. Peitsch, *Pac. Symp. Biocomput.* **1998**, 425 (1998).