A UNIFIED APPROACH TO THE PREDICTION OF PROTEIN STRUCTURE AND FUNCTION

JEFFREY SKOLNICK

Laboratory of Computational Genomics, Danforth Plant Science Center,
Creve Coeur, MO, U.S.A.

ANDRZEJ KOLINSKI

Laboratory of Computational Genomics, Danforth Plant Science Center,
Creve Coeur, MO, U.S.A.; and Department of Chemistry,
University of Warsaw, Warsaw, Poland

CONTENTS

Computational Methods for Protein Folding: A Special Volume of Advances in Chemical Physics, Volume 120, Edited by Richard A. Friesner. Series Editors I. Prigogine and Stuart A. Rice.
ISBN 0-471-20955-4. © 2002 John Wiley & Sons, Inc.

I. INTRODUCTION

In this postgenomic era, a key challenge is to interpret the information provided by the knowledge of the proteome, the set of protein sequences found in a given organism. Unfortunately, having a list of protein sequences in and of itself provides little insight; the key question is, What is the function of all of the proteins? Function covers many levels, ranging from molecular to cellular or physiological to phenotypical. By employing sequence-based methods that exploit evolutionary information, between 40% and 60% of the open reading frames (ORFs) in a given genome can be assigned some aspect of function ranging from physiological to biochemical function. Indeed, because of their considerable success, sequence alignment methods such as PSI-BLAST [1,2] and sequence motif (that is, local sequence descriptors) methods such as Prosite [3], Blocks [4], Prints [5,6], and Emotif [7] set the standard against which all

alternative approaches must be measured. However, sequence-based approaches increasingly fail as the protein families become more diverse [8]. The remaining unassigned ORFs, termed ORFans, represent an important challenge and represent an area where structure-based approaches to function prediction can play a significant role. One structure-based method combines one-dimensional information about sequence and structure and has had some success [9]. An alternative structure-based approach to function prediction that employs the sequence–structure–function paradigm has recently been developed [8,10–15]. Here, low-resolution models predicted by threading or *ab initio* folding are screened for matches to known active sites; if a match is found, then a functional assignment is made. However, this method requires a predicted structure of appropriate resolution. Structure prediction techniques will also play an important role in probe selection in structural genomics, where the ultimate goal is to experimentally determine the structure of all possible protein folds such that any newly found sequence is within modeling distance of an already solved structure. Thus, in this review, we examine the status of contemporary structure prediction approaches and demonstrate that the resulting (quite often low-resolution) models can be used both to identify the biochemical function of the protein and to dock known ligands to the correct binding sites.

Presently, there exist three approaches to protein structure prediction: homology modeling, threading, and *ab initio* folding. In homology modeling, the probe and template sequences are clearly evolutionarily related, and the structures of the probe and template are quite close to each other. The second structure prediction method is threading, where one attempts to find the closest matching structure in a library of already solved structures but where the structures can be analogous; that is, the two proteins are not necessarily evolutionarily related, but they adopt very similar structures. Ideally, threading should extend sequence-based approaches. Threading and homology modeling suffer from the fundamental disadvantage that an example of the fold of the sequence of interest must already have been solved in order for the method to be successful. Finally, there is *ab initio* folding where one attempts to fold a protein from a random conformation; obviously this is the hardest of the three methods of structure prediction, but it has the advantage that an example of the fold need not have been seen before. As detailed in what follows, a number of variants of *ab initio* folding use extensive information from threading. Such information might include local secondary structure information, supersecondary structure information, and/or predicted tertiary contacts. Indeed, the major focus of this review is to describe a unified approach to protein structure prediction that reduces to threading plus structure refinement when an example of the probe sequence is found; but if not, it incorporates information from weakly significant probe sequence–template structure matches and then does *ab initio* folding with the structural information gleaned from such matches. It has the advantage that it can predict a novel fold even though some of the information comes from threading on already solved structures.

II. OVERVIEW AND HISTORICAL PERSPECTIVE

A. Comparative Modeling Methods

Comparative modeling can be used to build the structure of those proteins whose sequence identity is above 30% or so with a protein template structure [16]. This usually consists of three steps: (1) Search for sequence similarity to a member of a set of carefully selected sequences with known three-dimensional structure; (2) use the detected structural template to build a molecular model; and (3) carefully validate the resulting models. In the recent CASP3 prediction experiment [17], encouraging results were reported by Bates and Sternberg [18], Blundell and coworkers [19], Yang and Honig [20], Dunbrack [21], and Fischer [22]. While the automated approach of Sali's MODELLER [23,24] did not do as well as others, it is nevertheless a widely used comparative modeling package. The results of CASP3 suggest that the key to a good model is to generate the best possible initial sequence alignment and to modify it as little as possible [25,26]. Thus, as the sequence identity of the probe and template moves into the twilight zone, the sequence alignments degrade with a comparable degradation in the quality of the model structures.

As an example of genome-scale comparative modeling using standard sequence alignment algorithms and MODELLER, Sanchez and Sali [27] recently scanned a portion of the yeast genome, *S. cerevisiae* [28]. They found homologous proteins of known structure for about 17% of the proteins (1071 sequences), and they built three-dimensional models for these yeast proteins. Only 40 of these modeled proteins had a previously determined experimental structure, and 236 proteins were related to a protein of known structure for the first time.

An obvious limitation of the above approach is that it requires a homologous protein whose structure is known. Depending on the genome, 15–25% of all sequences now have a homologous protein of known structure [29]. This percentage is slowly increasing as new structures are being solved at an increasing rate. Interestingly, the majority of newly solved structures exhibit an already known fold. At this point, it is still uncertain whether this indicates that proteins can adopt a limited number of folds or if it simply indicates a bias toward certain types of protein folds that crystallize relatively readily.

B. Threading

Threading is another means of predicting the tertiary structure of proteins. Here, for the sequence of interest, one attempts to find the closest matching structure in a library of known folds [30,31]. The paradigm of homology modeling is still followed with its three steps: (1) identifying the structural template, (2) creating the alignment, and (3) building the model. Thus, threading has limitations that are similar to classical homology modeling. First and foremost, an example of the correct structure must exist in the structural database that is being screened. If not, the method will fail. Second, the quality of the model is limited by the extent of actual structural similarity between the template and the probe structure. Until recently [32], one could not readjust the template structure to more correctly accommodate the probe sequence. While the quality of alignments generated by threading algorithms improved from CASP1 to CASP3 [17], it nevertheless remains problematic. Another question is whether threading recognizes distant homologies (i.e., a protein that is evolutionarily distant but still related to the template protein) as opposed to pure fold recognition targets (where the two proteins are evolutionarily unrelated, but have converged to the same fold). We note that for sequences that are evolutionarily very distant, convergent versus divergent evolution is very difficult to prove. Nevertheless, we still have the problem of identifying two proteins as having the same fold, when only about 65% of their sequences share a common core, with the possibility that the remainder of the fold differs significantly.

Next, we describe the features of existing threading algorithms that performed well in CASP3 as well as in the intervening period prior to CASP4. In the construction of a threading algorithm, one is faced with three choices: the type of energy used to assess the probe sequence–template structure suitability, the degree of detail used to describe interaction centers if multibody interactions are included, and the conformational search scheme employed to find the optimal sequence-structure alignment. In what follows, we address each of these three features in turn.

The first step in constructing a threading algorithm involves the choice of the potential used to describe the sequence-structure fitness and the potential for scoring functions containing more than one term; weights must be established. Among the kinds of energy terms that have been previously considered are the burial status of residues, secondary structure propensities and/or predicted secondary structure, additional penalty terms [33,34] (for example, those that compensate for different protein lengths), and the inclusion of pair or higher-order interactions between side chains. Contemporary algorithms often include an evolutionary component related to the sequence similarity between the template and the probe sequence [35]. Inclusion of such sequence-based terms improves the ability of the algorithm to recognize the correct structural template as well as the quality of the predicted alignment in the structural template [34, 36–39]. While such terms should not be needed in a structure-based approach, in practice they are found to be quite important.

If pair interactions are included, then the interaction centers must be selected, with common choices being the Cαs [40,41], the Cβs [42,43], the side-chain centers of mass, specially defined interaction centers [30,44], or any side-chain atom [45]. This defines the protein representation. Then, one must again choose the form of the interaction. Contact potentials [45,46], continuous distance-dependent potentials [42,47], and interaction environments [48] are the choices that have been made for the functional form of the pair energy.

Third, given an energy function, the optimal alignment between the probe sequence and each structural template must be found. Dynamic programming [49] is the best choice when local interaction schemes are used (e.g., when the energy consists of mutation matrices and secondary structure propensities). The situation when a nonlocal scoring function is used (e.g., pair interactions) is not as straightforward. Here, the problem is to update the interactions in the template structure to include the actual partners present in the probe sequence. To retain speed (a crucial feature if entire genomes are to be scanned), some workers employ dynamic programming with the "frozen" approximation (where the interaction partners or a set of local environmental preferences are taken from the template protein in the first threading pass) [45,50]. Iterative updating might follow this [45,48,51]. Still others employ double dynamic programming, which updates a subset of interactions recognized as being the most important in the first pass of the dynamic programming algorithm [42]. Other, more computationally intensive approaches evaluate the nonlocal scoring function directly and search for the optimal probe–template alignment by Monte Carlo [44] or branch-and-bound search strategies [30]. These have the advantage that the correct energy is evaluated, but unfortunately they are very CPU-intensive.

A problem with almost all threading search protocols is that they do not allow the actual template structure to adjust to reflect the actual structural modifications relative to the template structure that are actually present in the native conformation of the probe. For example, Monte Carlo and branch-and-bound strategies allow the partner from the probe sequence provided by the current probe–template alignment to be used, but they do not allow the template's backbone structure to readjust to accommodate the probe sequence. Such structural modifications should be quite important when the probe and template structure are analogous. As a simple example, when the probe's TYR replaces a GLY in the template protein, then the contacts associated with the amino acid at that position in the structure would be radically different. Yet, this effect is not accounted for at all in threading. However, the potential ability to recognize analogous structures is precisely the realm where threading should be the most valuable as compared to pure sequence-based methods.

As indicated above, because threading uses structure, it should be superior to sequence-based approaches that are one-dimensional and that assess the evolutionary relationship between sequences and thereby, by inference, their structural relationship. In practice, however, many of the most successful

fold-recognition approaches in CASP3 were pseudo one-dimensional and used evolutionary information that contributed a significant fraction of the selectivity [52] (typically implemented in the form of sequence profiles) plus predicted secondary structure. In particular, the Jones [53] and the Koretke groups [39] employed this type of approach, where secondary structure played an ancillary role. The Nishikawa group [54] also employed a hierarchy of local scoring functions to describe hydration, secondary structure, hydrogen bonding, and side-chain packing.

There were other successful approaches in CASP3 where structure played a more prominent role. For example, the Sippl group [55] employed burial energy and the frozen approximation to evaluate pair interactions, but unlike many others, they used a single sequence rather than sequence profiles or other implementations of multiple sequence information. While the Sippl approach is more structure-based, in order for dynamic programming to be used all interactions were made pseudo one-dimensional. The Bryant group [56] was unique in that they explicitly treated pair interactions within a structural core identified from the evolutionary conservation of structure across each protein family. In order for the core to be identified, a number of structures in the protein family must be solved. While this approach embodies the original idea of threading, they too employ a PSI-BLAST sequence-profile component. Indeed, they conclude that the combination of both sequence profiles and contact potentials improves the success rate relative to that when either of the terms is used alone. Because the Bryant group employs a nonlocal scoring function that *a priori* precludes dynamic programming, a Monte Carlo search procedure was used to find the best sequence–structure fitness. Unfortunately, these calculations are very CPU-intensive, thereby precluding the application of this approach on a genomic scale unless there are very substantial computer resources.

The general consensus was that CASP3 saw some progress in threading, with alignment quality improving from CASP2 [17,26,52], but, as pointed out by Murzin [52], threading "performs better on distant homology recognition targets than on 'pure' folding recognition targets. This bias probably resulted from the implementation of 'distant homology' filters." Thus, techniques that extend the ability of threading techniques to address "pure" fold recognition situations are still required. But, as Bryant and co-workers [35] have pointed out, the best results are found when a sequence–profile term is combined with threading potentials. These observations motivated the development of a new threading algorithm, PROSPECTOR (*PRO*tein *S*tructure *P*redictor *E*mploying *C*ombined *T*hreading to *O*ptimize *R*esults) [57], where it was demonstrated that pair interactions could significantly improve the sequence–structure specificity over that when only sequence–profile terms are used. However, when multiple scoring functions are combined, the resulting recognition ability is even larger. In Section IV, we discuss the results of this new approach in some detail, because it is a key component of a recently developed unified approach to protein structure prediction. But here we note that while considerable progress has been made in threading by a number of workers, we will have to await the results of CASP4 to assess the full extent of this progress as well as the limitations of such approaches.

C. *Ab Initio* Protein Structure Prediction

Due to the time scale of the protein folding process, which takes from milliseconds to minutes, at present, it is rather impractical to attempt protein structure assembly using all-atom detailed models. Indeed, contemporary computers allow classical molecular dynamics simulations of a protein surrounded by an appropriate number of water molecules over a much shorter period of time, corresponding to tens or hundreds of nanoseconds (depending on protein size). This inability to routinely access longer time scales stimulated numerous attempts to simplify the problem by reducing the number of explicitly treated degrees of freedom of the polypeptide chain and by simplifying the model of intra and intermolecular interactions. Such a reduction of the number of degrees of freedom could be achieved by assuming a united-atom representation of entire amino acid residues, by assuming a single-atom representation of the main chain and a similar representation of the side groups. The internal degrees of freedom of the side groups were frequently ignored in such models or were treated in an approximate fashion. Such a simplified protein representation also led to simplifications in the interaction scheme; for example, all reduced models either ignored the effect of water or implicitly treated it.

The first attempts at the reduced modeling of protein folding were undertaken about 25 years ago. In their classical work, Levitt and Warshel [58] proposed a model that later inspired other analogous simplifications of protein representation. They assumed two centers of interaction per residue, one associated with the alpha carbon and the second with the center of mass of the side group. There was a single degree of freedom per amino acid—the rotation around the Cα–Cα virtual bond—while the planar angle for the Cα trace was assumed to be constant [59]. A knowledge-based potential controlled the short-range interactions, while the interactions between the side groups were in the form of a Lennard-Jones potential (partially corrected for the hydrophobic effect). The sampling was done by means of classical molecular dynamics. Simulations of a small protein bovine pancreatic trypsin inhibitor sometimes produced structures resembling the native fold. The best structures had a root-mean-square-deviation (RMSD), from native in the range of 6.5 Å. Later, Kuntz et al. [60,61], Hagler and Honig [62], and Wilson and Doniach [63] studied somewhat similar continuous models. The results were of comparable quality; some qualitative features of small protein folds were sometimes recovered in their simulations.

More recently, continuous-space models with more structural details were proposed and investigated with respect to their ability to predict the native conformation of a protein. Sun [64] examined models with an all-atom representation of the main chain and a single united atom representation of the side groups. Knowledge-based statistical potentials described the interactions between the side groups. Interestingly, his study demonstrated that a genetic algorithm could quite efficiently sample the conformational space of the chain. For small peptides (mellitin, pancreatic polypeptide inhibitor, and apamin), proper structures were predicted whose accuracy ranged from 1.66 Å to 4.5 Å, depending on peptide size. A similar model, but with two united atoms per side chain (for the larger amino acids), was studied by Wallqvist and Ullner [65]. Results for pancreatic polypeptide inhibitor were slightly more accurate, probably due to the better packing of the model side chains. Such reduced continuous models were explored not only as a means of protein structure prediction but also as a tool for investigating the general aspects of protein folding dynamics and thermodynamics [66,67].

Pedersen and Moult [68] proposed a very interesting approach to protein structure prediction. They assumed an all-heavy atom representation of the protein with knowledge-based potentials describing intraprotein interactions. As a sampling method, they used a combination of Monte Carlo (MC) and genetic algorithms. The MC runs produced a set of structures for the starting population of the genetic algorithm (GA). The crossover points were selected in the regions of the largest structural flexibility, as detected during the MC runs. MC simulations were also performed between crossover events in the GA scheme. Low- to moderate-resolution protein fragments and the approximate folds of small proteins have been successfully predicted by this method. Unfortunately, it appears that the applicability of this method is limited to rather small proteins.

Even reduced models of proteins have a large number of conformational degrees of freedom, and an effective sampling of the long-time processes for larger proteins in a continuous space could be very difficult if not impossible. To further simplify the problem, discrete or lattice models were proposed and examined. Early studies of the lattice proteins focused not on structure prediction but rather on understanding the fundamentals of protein folding thermodynamics and some aspects of the folding dynamics. These works were pioneered by Gō et al. [69], and then followed by Krigbaum and Lin [70,71], Skolnick and Kolinski [72–84], Sikorski and Skolnick [85–88], Chan and Dill [89–92], Dill et al. [93–96], Sali et al. [97,98], Shakhnovich et al. [99–105], and others [106–111]. Since the subject of this chapter is protein structure prediction and due to the existence of excellent reviews on the subject, we refrain from a more detailed review of these works.

Probably the first attempt to predict the native structure of a protein in an *ab initio* fashion within the framework of a lattice representation is due to

Dashevskii [112]. A diamond lattice chain was used to approximate the polypeptide conformations. A chain growth algorithm executed the sampling of conformational space. Compact structures resembling native folds of small polypeptides were generated and identified by a simple force field. Next, Covell investigated a simple cubic lattice model of real proteins [113]. The behavior was controlled by the force field that consisted entirely of long-range interactions that included a pairwise, knowledge-based potential, a surface term, and a potential that corrects the local packing of the model chain. The quality of crude folds generated by this method were not worse than the quality of folds obtained using early continuous models. Covell and Jernigan [114] studied five small globular proteins by the enumeration of all possible compact conformations of a body-centered cubic lattice chain. They found that the closest to native conformation could always be found within the top 2% of the lowest-energy structures, as assessed by a knowledge-based interaction scheme.

Hinds and Levitt [115] proposed an interesting lattice model of proteins. In a diamond lattice chain, a single lattice vertex represents several residues of a real protein. An elaborate statistical potential was employed to mimic the mean interactions between such defined protein segments. Frequently, correct folds of low resolution were generated among the compact structures enforced by the sampling scheme.

Kolinski and Skolnick [75–84,116–120] developed a series of high-coordination lattice models of globular proteins. Lattices of various resolution were employed to mimic the conformation of the Cα trace of real proteins, from three-dimensional "chess-knight"-type lattices to a high coordination lattice with 90 lattice vectors to represent possible orientations of the Cα–Cα virtual bonds. The models employed in the test structure predictions [118,121–123] had additional interaction centers to represent the side groups. For each side chain, a single-sphere, multiple rotamer representation was assumed. The force field of each of these models contained several terms mimicking the short-range interactions, explicitly cooperative hydrogen bonds, one body, and pairwise and multibody long-range interactions with an implicit averaged effect of the water molecules. It has been shown for several cases of small globular proteins [118] and simple multimeric molecular assemblies [124–126] that such models can generate correct low- to moderate-resolution (high-resolution in the case of leucine zippers) folds during Monte Carlo simulated annealing computer experiments.

Various recently developed methods for *ab initio* protein structure predictions were tested during the CASP3 (Critical Assessment of Techniques for Protein Structure Prediction) exercises, concluded in December 1998 in Asilomar, California [127]. A number of new techniques have been developed before that time, and a number of them constitute qualitative progress in *ab initio* prediction with respect to the previous CASPs (held every two years).

The ROSETTA method proposed by Baker and co-workers [128] is very innovative. The method consists of several steps. First, a multiple sequence alignment for a sequence of interest was prepared, and the secondary structure prediction is made using the PHD server based on Rost and Sander's [129–131] secondary prediction technique. Secondary structure predictions and sequence alignments were then used to extract the most plausible 3- to 9-residue structural fragments (25 fragments for each segment of the query sequence) from the structural database (according to the secondary structure prediction and the sequence similarity). Then a Monte Carlo algorithm employing a random insertion of fragments into the structure was used to build the three-dimensional structure. The scoring function contained a hydrophobic burial term, elements of electrostatics, a disulfide bond bias, and a sequence-independent term that evaluates the packing of secondary structure elements. The top 25 (of 1200 generated) structures frequently contained the proper fold. The best five structures exhibiting a single hydrophobic core were selected by "visual inspection." This could be considered to be a flaw of the method (at this stage of development). It would be difficult to do a manual evaluation of the predictions on a massive scale. Nevertheless, for 18 targets, four predictions were globally correct (with an RMSD range of 4–6 Å for the native structure), and the majority of their predictions contained significant fragments of structure that were correct. It should be noted that a somewhat similar idea of protein structure assembly using predefined fragments and the Monte Carlo method was also pursued in the method developed by Jones [132] and tested during the CASP2 exercise.

A number of other groups made good predictions on a fraction of difficult *ab initio* target proteins. Ortiz et al. [133] applied a high coordination lattice model developed by Kolinski and Skolnick [122,123] to a number of small target proteins. Monte Carlo simulated annealing calculations started from random expanded conformations of the target proteins. The model assumed a 90-basis vector representation of the alpha carbon trace that has a 1.2 Å resolution due to the spacing of the underlying cubic lattice grid. Off-lattice single-sphere side chains could assume multiple orientations with respect to the backbone, thereby mimicking the distribution of rotamers for particular amino acids. The generic force field of the model consisted of knowledge-based potentials (derived from the statistics of the regularities seen in known protein structures) for short-range interactions, one body burial, pairwise and multibody surface long-range interactions, and terms simulating the regularity and cooperativity of the main-chain hydrogen bond network. Additionally, a weak bias toward predicted secondary structure (obtained from multiple sequence alignments + secondary structure prediction from PHD [129–131]) and weak theoretically predicted long-range contact restraints from correlated mutation analysis were implemented in the interaction scheme [134–138]. Contact prediction was based on the analysis of correlated mutations in sequences detected by multiple sequence alignments. For some targets, the globally correct fold or large fragments of the structure were correctly predicted. The method was capable of assembling low-resolution novel folds. The level of success during the CASP3 exercise was on the same level as reported for test predictions made for a series of small globular proteins prior to CASP3 [137].

A similar methodology, but one based on a completely different protein representation [139,140] (that are discussed in Sections V and VI), was employed by Kolinski and co-workers with a similar fraction of correctly predicted structures [133]. An important advantage of this method was its computational speed and nicer scaling of computational cost against protein chain length. Thus, the prediction of structures of larger proteins via *ab initio* folding became possible.

Osguthorpe [141] employed a continuous model and molecular dynamics simulated annealing. In spite of the use of a quite detailed model (main chain united atoms and up to three united atoms per residue), its very flexible chain geometry enabled efficient sampling. The potentials were derived from the statistics of known protein structures. The method enabled us to obtain correct predictions of substantial fractions of the structure of the attempted targets, and for one of the difficult targets, the prediction resulting from this method was the most accurate.

A very interesting hierarchical procedure has been used by Samudrala et al. [142]. First, as previously proposed by Hinds and Levitt [143], all compact conformations of test proteins were enumerated using the diamond lattice model with multiple residues per chain unit. The best (according to the force field of the lattice model) structures were then selected for further consideration. Subsequently, the all-atom structures were reconstructed by fitting the predicted secondary structure fragments to the lattice models. These structures were subject to energy minimization using an all-atom force field and spatial restraints of the lattice models. The optimized structures were scored by a combination of all-atom and residue-based knowledge-based potentials [144]. Then, distance geometry [145] was used to generate a number of possible "consensus" models. The local geometry of predicted secondary structure was again fitted to the resulting models. Finally, the resulting all-atom models were optimized and rank-ordered according to energy. A number of qualitatively correct protein fragments of significant size were correctly predicted. The method appears to be very robust and (as pointed out by the authors) it was likely that it could be further improved. Probably the major weakness of the method in its present form is in the small fraction of good structures in the initial pool of lattice models.

The method developed by Scheraga and co-workers [146] and used in CASP3 is based on the global optimization of the potential energy of a united atom model [147]. Due to the force-field design of the model, which is based on basic physical principles, this method is very close to a purely thermodynamic approach. In this respect, it qualitatively differs from the previously outlined methods. This off-lattice protein model has a united atom representation of the alpha carbons, side groups, and peptide bond group, with fixed bond lengths and variable bond angles. The interaction potentials between united atoms describe the mean free energy of interactions and account in an implicit way for the average solvent effect and cooperativity of the hydrogen bonds [148]. The optimization is performed by means of the Conformational Space Annealing technique [147], which subsequently narrows the search regions and finally finds distinct families of low-energy conformations. The lowest-energy, reduced model conformations are subsequently converted into the all-atom models and optimized by electrostatically driven Monte Carlo simulations [149]. For a fraction of CASP3 targets, this method produced exceptionally good predictions. The method seems to perform much better on helical proteins than on β or α/β proteins.

D. Choice of Sampling Scheme

In the past, different methods of sampling of protein model conformational space have been employed with various degrees of success. Traditional molecular dynamics can be used only in the case of continuous models. Other sampling schemes, including a variety of Monte Carlo methods, genetic algorithms, and combinations of these methods, could be applied to continuous as well as to the discrete (including lattice representation) models.

In general, the choice of the simulation/optimization algorithm depends on the aim of the studies. Different procedures are needed for the study of protein dynamics and folding pathways from those procedures that are just targeted to find the lowest-energy conformations of model polypeptides.

Monte Carlo procedures for chain molecules [150] use a wide spectrum of strategies for conformational updating. In some algorithms, the updates are global, as in the chain growth algorithms, whereas other algorithms employ pivot moves of a large part of the model chain. In other algorithms, the trial modifications are local, involving only a small portion of the chain or a small distance displacement of a larger part of the chain. Sometimes, the local and global modifications were combined in the same algorithm.

What is the relationship between the molecular dynamics simulations of a continuous model and an isothermal Monte Carlo trajectory of an otherwise similar discretized (or lattice) model? When only local (and small distance) moves are applied in a properly controlled random (or rather pseudorandom) scheme, the discrete models mimic the coarse-grained Brownian dynamics of the chain. The Monte Carlo trajectory could be then interpreted as the numerical solution to a stochastic equation of motion. Of course, the short-time dynamics (the time scale of a single elementary move in the Monte Carlo scheme) of the discrete model has no physical meaning. However, the long-time dynamics should be qualitatively correct, albeit with possible distortions of the time scale of various dynamic events. Such an equivalence of the molecular dynamics and stochastic dynamics of equivalent off-lattice and lattice-simplified protein models has been demonstrated in the past by Rey and Skolnick [151], and by Skolnick and Kolinski [152]. Recent studies have shown that Monte Carlo folding pathways observed for high-coordination lattice models reproduce the qualitative picture of folding dynamics seen in experiments [153]. Thus, it could be rather safely assumed that Monte Carlo lattice dynamics can be used in meaningful studies of protein dynamics, folding pathways, the mechanism of multimeric protein assembly and other aspects of biopolymer dynamics. The validity of protein dynamics studies using discrete models depends more on the assumed accuracy of the protein representation and its force field than on the particular sampling scheme. However, some oversimplified discrete models may face serious ergodicity problems. This aspect of Monte Carlo simulations always needs to be carefully examined.

Isothermal simulations (molecular dynamics or Monte Carlo) provide characteristics of the system's properties at a single temperature. Numerous simulations at various temperatures (above and below the folding transition temperature) are needed to gain some insight into the thermodynamics of the folding process. There is a very serious problem associated with the extremely slow relaxation of protein models in the dense globular state. The local barriers in the energy landscape near the folded state are high and the sampling becomes ineffective. Thus the computer studies employing straightforward MD or canonical MC algorithms became prohibitively expensive. Essentially, the same applies to various simulated annealing strategies. In all cases, the design of sampling details could be very important. For example, properly designed local moves can "jump over" the high local energy barriers, thereby speeding up the sampling of the entire conformational space.

Mulicanonical [154] (or entropy sampling Monte Carlo [108–110]) simulations provide more complete data on folding thermodynamics [116,155–157]. Due to their differently defined transition probabilities in the sampling scheme, energy barriers became much less important, but are substituted by entropic barriers. From a single series of simulations, it is possible to obtain an estimation of all thermodynamic functions (energy, free energy, and entropy) over a wide range of temperatures. However, the cost of such computations grows rapidly with the system size and its complexity.

A somewhat simpler, but by no means trivial, task is to find the lowest energy state of the model polypeptide. Due to the thermodynamic hypothesis [158], which postulates that native proteins are in the global minimum of the conformational energy, the minimum energy state of a properly designed protein model should closely mimic the folded conformation. A variety of strategies have been developed to solve this global minimum problem [159]. For a relatively simple system, when the total energy could be expressed in the analytical form, it is possible to solve the problem in a deterministic fashion [160]. For more complex (i.e., realistic models of proteins) systems, existing methods do not guarantee that the lowest energy conformation will be found. The number of possible conformations and the rugged energy landscape make a systematic search impractical.

Simulated annealing, ESMC [108,109,161], Monte Carlo with minimization [162], genetic algorithms [64,163–165], and the combination of genetic algorithms with Monte Carlo sampling have been successfully used in the past to find the near-native conformations of reduced models of small proteins [68].

Recently, a number of studies have focused on the comparison of various Monte Carlo strategies for finding the global minimum of a protein model [166–168]. Probably the most straightforward of these search strategies is simulated annealing, where the system temperature is gradually lowered during the simulations, starting from a relatively high temperature (above the folding transition) and ending at a low temperature below the folding temperature (usually well below due to thermal fluctuations). When on repeated runs starting from different initial states, the same conformation is recovered; one may assume that there is a good chance that the global minimum has indeed been found. However, for difficult problems, simulated annealing runs (or at least a substantial fraction of the runs) could be trapped in local energy minima. Some of the local minima could be close to the model's representation of the native state, whereas others could correspond to conformations that are far away from the properly folded state. There is no simple test of convergence in the simulated annealing method. The efficiency of the simulated annealing method could be considerably improved by a certain modification of transition acceptance criteria. For instance, one may perform local minimization before and after the transition and then apply the Metropolis criterion to the locally lowest energy pairs or conformations [16]. This way, the sampling procedure can avoid visits to a large fraction of irrelevant local energy minima.

In contrast to simulated annealing, sampling techniques within the multicanonical ensemble have some internal convergence tests. In a version of this technique, called entropy sampling Monte Carlo [108–110], the estimation of the system's entropy is built by a sampling process that is controlled by the density of states of particular discretized levels of conformational energy. When converged, all energy levels, including the lowest energy, should be sampled with the same frequency. The ESMC method is "quasi-deterministic": The data from the preceding simulations could be used to improve the accuracy in the successive runs. In principle, when converged, ESMC should find the lowest energy state. In practice, the energy spectrum near the lowest energy state could be associated with large entropy barriers, and the lowest energy state could be not detected in spite of the apparent convergence—that is a constant density of visited states in the remaining low-energy portion of the energy spectrum. The rate of convergence of the ESMC method into the low-energy portion of the energy landscape could be accelerated by the artificial deformation of the entropy curve (artificial increase of the density of states) in the less important, high-energy range [156].

The replica exchange Monte Carlo method [169] addresses the problem of local minima in a different way. A number of copies of the model system are simulated by means of a standard Metropolis scheme at various temperatures. The temperature range covers temperatures from a temperature well above the folding temperature down to a temperature below the folding transition temperature. Occasionally, the replicas are randomly swapped according to a criterion that depends on temperature difference and the energy difference. Thus, the low-energy conformations at a higher temperature have a chance to be moved to a lower temperature. As a result, the copies of the system sample not only the conformational space but also move between various temperatures. At high temperatures, the energy barriers could be surmounted easily; at low temperatures the vicinities of energy landscape "valleys" are efficiently sampled.

Comparison of the computational cost of finding the lowest energy state for a simple protein-like copolymer model [168] shows that replica exchange Monte Carlo (REMC) is much more efficient than simple Metropolis sampling with a simulated annealing protocol in spite of the fact that multiple copies of the system have to be simulated. The REMC method also finds the low-energy conformations many times faster than the ESMC method. Thus, it appears that the REMC method (or its variants) could be a method of choice for use in the *ab initio* folding of reduced protein models, where finding the lowest energy state is the main goal of computational experiment. Due to the very efficient sampling by the REMC method, the samples at various temperatures could be used for the "umbrella"-type estimation of the system entropy. That may extend the applications of the REMC method into cost-efficient studies of protein folding thermodynamics.

III. OVERVIEW OF THE UNIFIED FOLDING METHOD

When faced with the problem of predicting the tertiary structure of an unknown sequence, one typically runs PSI-BLAST [170] over sequences from the structures in the protein data bank [171]. Then, if this does not work, one runs a threading program to see if it detects a significant probe–template match. Even if either of these two cases is successful, for nontrivial cases often the alignments of the probe sequence may be in error, and there may be gaps in the alignment of the probe sequence to the template structure and/or sometimes there are long unaligned regions. If both methods fail, then *ab initio* folding is the requisite

structure prediction method. Thus, ideally one would like to have a unified approach that automatically treats these possibilities. In what follows, we describe one recently developed unified approach.

An overview of the idea is given in Fig. 1. First, one runs our threading algorithm, PROSPECTOR [57], and establishes if there is a significant probe sequence–template structure match. If so, the template is used as a soft bias in a generalized comparative modeling approach that involves *ab initio* folding in the vicinity of the template in a reduced protein model. Threading also provides predicted secondary structure and tertiary contacts that are not restricted to the template structure but can be extracted from other structures. This allows the possibility of fold prediction in those regions absent in the alignment of the probe sequence to the template structure. The advantage of this generalized comparative modeling is that it can improve the initial alignment generated by the threading algorithm and can provide a structure prediction for the unaligned

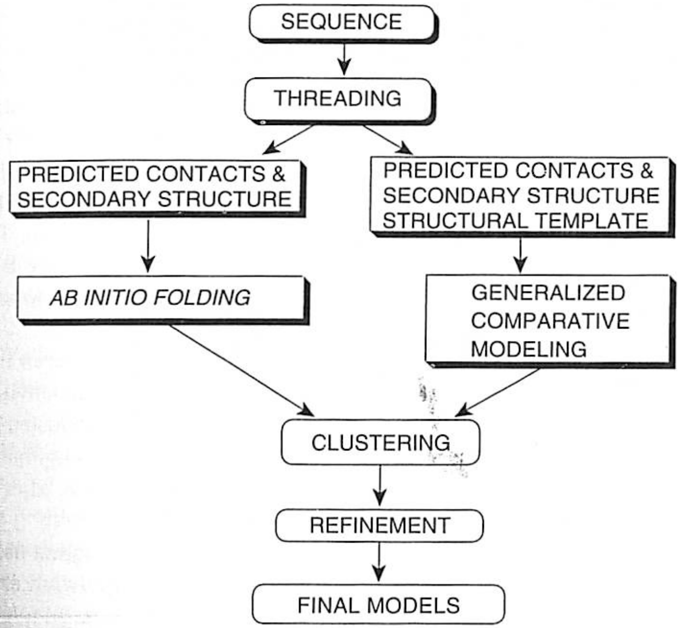

SEQUENCE
THREADING
PREDICTED CONTACTS & SECONDARY STRUCTURE
PREDICTED CONTACTS & SECONDARY STRUCTURE STRUCTURAL TEMPLATE
AB INITIO FOLDING
GENERALIZED COMPARATIVE MODELING
CLUSTERING
REFINEMENT
FINAL MODELS

Figure 1. Flow chart describing the unified approach to protein structure prediction. First, threading is done. If a significant hit to a template is found, then generalized comparative modeling in the vicinity of the template but supplemented by predicted secondary structure and contacts possibly from other templates is done. If no significant probe sequence–template structure match is found, then consensus contacts and sets of local distances in the top 20 scoring structures are extracted and employed as restraints in an *ab initio* folding algorithm. Once a sufficient number of simulations (typically 100) are done, the structures are clustered, full atomic models are built in the refinement step; and then using a new, distant-dependent atomic pair potential [204], the top five scoring structures are selected.

regions of the probe sequence. On the other hand, if there is no significant match to a template, then the predicted secondary structure and tertiary contacts extracted from threading are passed to an *ab initio* folding algorithm that uses the same reduced protein model. Then, for both generalized comparative modeling and *ab initio* folding, the resulting structures are clustered, atomic detail is added and the results are reported.

IV. THREADING RESULTS

A. First-Pass Threading

Recently, to build on the strengths and address the weaknesses of existing threading approaches, we have developed a new threading algorithm called PROSPECTOR (*PRO*tein *S*tructure *P*redictor *E*mploying *C*ombined *T*hreading to *O*ptimize *R*esults) [57], which runs sufficiently quickly so that entire genomes can be scanned in the matter of several days on a standard workstation or PC. During the course of the development of this program, we noticed that sequence profiles generated from the BLOSUM 62 matrix [172] often generated reasonable alignments between the probe and template sequences, even when the alignment score was insignificant. This suggested that the first stage of a hierarchical approach to threading should employ a sequence-profile [170,173, 174] (using a sequence profile plus a three-state secondary structure prediction scheme gave worse results) to generate the initial probe sequence to template structure alignment. We call this the "partly thawed" approximation. Then, the resulting alignment of the probe sequence in the template structure is used to calculate the partners for the evaluation of the pair interactions. Previously, in the first iteration of the frozen approximation [45], the partners were taken from the template structure. This worked well only when the environments in the probe and template structures were similar, but more often than not the environments were quite different. On successive iterations, in the so-called defrosted approximation [45] where the partners were taken from the previous alignment, there were times when the resulting algorithm never converged. Here, after the first initial alignment, quite good results were obtained.

The database for multiple sequence alignment (MSA) generation used in the construction of the sequence profile combines Swissprot (http://www.expasy.ch/sprot/) and the genome sequence database (ftp://kegg.genome.ad.jp/genomes/genes). First, a profile for relatively closely related sequences, whose sequence identity lies between 35% and 90%, is calculated. These sequences are selected from the composite database by FASTA [175,176]. Then, pairwise sequence alignments with the probe sequence are generated using CLUSTALW [177], and a sequence profile is generated. We term this the "closely" related set of alignments. To this set, we add additional sequences whose *E* value in FASTA is less than 10, use CLUSTALW to generate pairwise alignments, and then generate a profile for distantly related sequences; these are termed the "distantly" related set of alignments. The goal here is to have two sequence profiles: one that is more sensitive to more closely related sequences and another that can sometimes detect more distantly related sequences.

The first step of the threading protocol is to independently scan the structural database of interest using each of the sequence-profiles with a Needleman–Wunsch type of global alignment program [49]. Each of the two sequence profiles generates an alignment of the sequence in each of the template structures. Each alignment is used to identify the partners in the probe sequence to be used in the calculation of the pair interactions. Here we use our previously developed side-chain contact potential averaged over all homologs which includes a contribution from contacting fragments that have weak sequence similarity to each member of the close set of probe sequences [178]. Furthermore, we also use a pseudo energy term that describes the preferences for consecutive types of amino acids to adopt a given type of secondary structure. This secondary structure propensity term is also averaged over homologs, and thus it results in a secondary structure propensity profile. For each scoring function, close (distant) sequence profile, and close (distant) sequence plus pair interactions plus the secondary structure propensity profile, we scan the structural database and output the top five scoring structures. Thus, a total of 20 possible structures are output, along with their alignments.

B. Application to the Fischer Database

As a test case, we have focused on the Fischer database [179] that is comprised of 301 template structures and 68 probe sequences. We tried a variety of approaches on this database before deciding on the aforementioned combination of parameters. We just summarize the results of these studies here. For a given scoring function, the Needleman–Wunsch global alignment algorithm recognized more correct probe–template pairs than did the Smith–Waterman [180] local alignment algorithm. We also tried using the secondary structure profiles as the initial step in generating the probe–template alignment for pair evaluation. Secondary structure profiles alone only correctly recognize 18 cases in the first position, whereas secondary structure profiles plus pair profiles correctly assign 29 cases. This clear improvement shows the utility of pair potentials in this approach; nevertheless, even 29 recognized pairs is rather poor performance. The major improvement in fold recognition comes, as others have observed, when sequence profiles are used. Even if the sequence profile is turned off completely but is used to generate the alignment, the number of correctly recognized pairs increases to 35 correct probe–template pairs in the top position. In all cases, inclusion of pair interactions improves the yield of correct probe–template matches.

We summarize our results using PROSPECTOR1 in Table I (the first pass of PROSPECTOR). One of the best alternative methods is that of Gonnet, which

TABLE I
Summary of Threading Results on the Fischer Database for Different Scoring Functions[a]

Method	Number of Fischer Pairs in the First Position	Number of Fischer Pairs in the Top 5(4) Positions	Number of Fischer Pairs in the Top 10(8) Positions
	PROSPECTOR1		
"Close" sequence profile	44	46(46)	49(47)
"Close" sequence profile plus secondary structure plus pair profile	45	55(53)	56(55)
"Distant" sequence profile	46	53(51)	53(53)
"Distant" sequence profile plus secondary structure plus pair profile	52	56(56)	59(57)
Hierarchy of four scoring methods	59	63(62)	65(63)
Hierarchy of three scoring functions (as above but withoat the "distant" sequence-profiles)	58	62	64
	PROSPECTOR2		
"Close" PROSPECTOR2. sequence profile plus protein specific pair and secondary structure potentials profile	48	51(51)	58(58)
"Distant" sequence profile plus protein specific pair and secondary structure potentials	51	59(59)	59(59)
Hierarchy of four scoring methods	61	64(64)	65(65)
Hierarchy of three scoring functions (as above but without the "Distant" sequence profiles)	60	64	65
	Other Methods		
Simple Blast[1]	27	—	—
PSI-BLAST restricted to the Fischer database [170,182]	24	37(36)	40(39)
PSI-BLAST using extensive sequence database and PSSM constructured using IMPALA [247]	41	46(46)_	47(46)
Original GKS threading program [45]	22	30	34
Hybrid threading [181]	52	57	60
Best UCLA benchmark results as of 2/4/00 which is prediction of secondary structure plus mult-gonnet [34]	52	(56)	(58)

[a] Results are reported in both the top 5(4) and top 10(8) positions [181], with the number in parenthesis given by the UCLA benchmark website (http://www.doembi.ucla.edu/people/fischer/BENCH/table1.html).

recognizes 52 proteins in the top position, the same number as the distant profile plus pair interactions recognizes, but if a hierarchical method is used, then ours is clearly the best, because 59 proteins are recognized in the top position. It is clearly superior to all our early efforts as well as to the alternative hybrid method [181], BLAST [1], and PSI-BLAST [170,182]. It might be argued that because we use four scoring functions while the hybrid method uses only three, this is not a strictly fair comparison. If we eliminate those results obtained from the "distant" sequence profiles, then we obtain 58, 62, and 64 cases in the top 1, 5, and 10 position as compared to 52, 57, and 60, respectively, of Gonnet.

We then applied the method to a second Fischer benchmark comprised of 29 probe–template pairs and scanned each probe sequence against the original Fischer structural database plus an additional 19 template structures (http://www.doembi.ucla.edu/people/fischer/BENCH/tablepairs2.html). We have only been able to find 27 of the 29 probe sequences and have reported our results accordingly. PROSPECTOR1 places 17 correct pairs in the top position, and it also places 21 and 22 in the top four and eight positions, respectively. This is the same as the best reported results of 17 correctly identified pairs. However, in our case one probe, "stel," which is supposed to be matched to 2azaA, selects 2pcy in the top position, which has the same core as 2azaA. Then, we have 18, 19 (19), and 20 (20) correct matches in the top position and top five (four) and ten (eight) positions, respectively. Thus, we have somewhat better results than previous workers.

C. Iterative Threading

1. *General Idea*

Just as PSI-BLAST [170] can increase its specificity by iteration, so can threading. In fact, the set of structures selected by PROSPECTOR contains additional information even beyond providing for a structural match. If we look at the set of 20 structures that are selected as being the best scoring sequence–template structure pairs, it is possible to extract additional information by looking for consensus predictions. By way of illustration, we consider the prediction of tertiary contacts. We focus on all contacts between residues that are at least five residues apart, and we count the predicted contacts generated by the aligned regions of structure. If there is a consensus (i.e., at least three contacts are consistently predicted), then we employ this information in two ways: (1) to enhance the specificity of threading by constructing a protein-specific, threading-based pair potential and (2) as described in Section IV.F, to predict tertiary contacts.

Using a previously derived formalism to convert contacts into a pair potential [178], we derive a set of protein-specific potentials, where the contacts are not only extracted from fragments with weak sequence similarity, but rather are generated by consensus contacts in the threaded structures. We use the arithmetic average of this potential and the previous iteration's pair potential in the next iteration of threading. This case is termed the "close" and "distant" protein-specific potentials, and we call the threading method that employs these terms PROSPECTOR2.

2. *Application of PROSPECTOR2 to the Fischer Database*

The results from PROSPECTOR2 are also reported in Table I. The "close" case now recognizes 48 proteins as compared to 45 in the top position. The "distant" case recognizes 51 as compared to 52 previously, but the composite of the four scoring functions now recognizes 61, 64, and 65 proteins in the top position as compared to 59, 63, and 65 in the top, top five, and top ten positions, respectively, for PROSPECTOR2. In all cases, the method improves when pair potentials are used as compared to that when the corresponding sequence profile alone is used. Similarly for in the second Fischer database, a total of 17, 20, and 20 proteins are recognized in the top, top five, and top ten positions, respectively.

D. Genome-Scale Iterative Threading

In tests on genome scale threading, we found that the optimum number of correctly recognized folds was found on the third iteration, PROSPECTOR3. However, because of the computational cost of constructing pair potentials that used local sequence fragment similarity, in our preliminary study and in the interest of computational tractability we employed the best quasi-chemical pair scale [183]. We term this PROSPECTORQUASI1-3. Furthermore, to deal with the problem of very large proteins that may contain more than one domain, in addition to threading the entire sequence, we also threaded 150 residue fragments, starting at the first residue and then shifting by 25 residues until the final fragment of possibly shorter length is scanned. This allows for the detection of domains. For genome-scale threading, our structure library consists of 2466 sequences constructed so that no pair of proteins has greater than 35% sequence identity between them.

1. *M. genitalium*

This genome consists of 480 ORFs [184]. The first pass of PROSPECTOR, PROSPECTORQUASI1 assigns 153 proteins to a structure in the protein databank. The second pass, PROSPECTORQUASI2, assigns 182, and the third pass, PROSPECTORQUASI3, assigns 194. This constitutes an assignment of 40% of the genome. All assignments are made using an automated protocol based on the score significance. Of these 194 structural predictions, all but three are correct. In contrast, several years ago Fischer and Eisenberg [185] assigned the folds of 103 out of a total of 468 proteins by their threading algorithm. Gerstein has reported identification of 211 proteins using PSI-BLAST [186,187]. Genethreader assigns

200 proteins, but for 15 of them the assignment appears to be incorrect [188] as assessed by a consensus of Gerstein's results (http://bioinfo.mbb.yale.edu/genome/MG/) and PROSPECTORQUASI3.

2. *E. coli*

The *E. coli* genome contains 4289 ORFs [189], for which PROSPECTOR-QUASI3 assigns 1716 ORFs to structures in the Protein Data Bank. This constitutes about 40% of the genome. Interestingly, this is the same percentage of structures as was assigned in *M. genitalium*. In contrast, without the use of active site filters, a total of 1250 confident structure predictions have been made, using a sequence profile-based method [190].

E. Extension of PROSPECTOR to Include an Orientation-Dependent Pair Potential

To enhance specificity, we next replaced the pair potential by one that is orientation dependent and again perform three iterations of modified PRO-SPECTOR, PROSPECTORIEN1-3. In applications to the Fischer database, we found that, on average, PROSPECTORIEN3 generates the most accurate probe–template alignments. The resulting set of structures constitutes the initial model that will be subjected to the generalized comparative modeling described in Section V.

F. Threading-Based Prediction of Tertiary Contacts

For a given iteration, the set of 20 top-scoring structures can also be used to predict the tertiary contacts in the probe protein. Again we demand that a given pair of contacts occurs in at least 25% of the top-scoring structures. For each interaction of PROSPECTOR1-3 and PROSPECTORIEN1-3, we collect the predicted contacts. The sets of contacts are then pooled.

Next we report our results for the set of 18 small proteins that constituted part of the validation set for the MONSSTER *ab initio* folding algorithm [191]. Of course, in this 18-protein test set, care is taken to remove all homologous proteins to the probe sequence from our structural database, and all proteins whose global root-mean-square deviations (RMSD) from native that are less than 8.5 Å are also excluded. On average, 28% of the contacts are correct, and 69% are correct within two residues. The correlated mutation analysis gives, on average, 34% correct and 82% correct within ±2 residues [191–193]. While the threading-based method has somewhat lower accuracy, in contrast to the correlated mutation analysis, it can be readily automated. Note that a contact-prediction accuracy of about 70% correct within ±2 residues is sufficient for the successful assembly of the global fold using the MONSSTER *ab initio* structure prediction program [191,193].

TABLE II
Comparison of Contact Prediction Accuracy for CASP3 Targets for Threading and Correlated Mutation Based Approaches[a]

Name of Protein	Number of Contacts Predicted	δ = 0 From Threading	δ = 0 Mutation Analysis	δ = 2 From Threading	δ = 2 Mutation Analysis	δ = 3 From Threading	δ = 3 Mutation Analysis
1jwe_	16	0.19	0.14	0.5	0.44	0.5	0.65
1eh2_	22	0.68	0.14	0.91	0.73	0.91	0.98
1bqv_	19	0.05	0	0.53	0.13	0.53	0.5
1ck5B	22	0.14	0.02	0.59	0.4	0.55	0.51
Average		0.265	0.075	0.63	0.43	0.62	0.66

[a]% of contacts correct with δ = 1m1 residues of a correctly predicted contact.

Turning to the results of CASP3, the correlated mutation analysis performed considerably poorer, whereas threading-based contact prediction was better [133]. In Table II, for four of these proteins, we show the predicted contact results and compare them to correlated mutation analysis. Now, within ±2 residues, 63% of the contacts are correct as predicted by the threading-based method as compared to 43% from the correlated mutation analysis; this is a qualitatively significant improvement. Within ±3 residues, correlated mutation analysis is slightly more accurate at 66% versus 62% from the threading-based contact predictions. Here again, we excluded all analogous and homologous proteins in the prediction of contacts from the analysis of consensus contacts in the alignments generated by PROSPECTOR1-3 and PROSPECTORORIEN1-3.

In Table III we present the set of predicted contact results for 28 proteins that that will be subject to *ab initio* folding in Section VI. Again the requisite contact prediction accuracy is achieved, with 31% of the contacts exactly predicted on average and 73% correctly predicted on average within ±2 residues. If we use the threshold of 70% prediction accuracy as indicative that the folding simulation will be successful, then, as shown in Section IV, 20 of these 28 proteins should be foldable. The asterisk indicates those proteins that are foldable, as assessed by the presence of a cluster of structures whose RMSD from native is less than 6.5 Å. In practice, of the 28 proteins, 13 are foldable. In addition, another two whose contact prediction accuracy is less than 70% correct within ±2 residues are also foldable. Of course, the presence of reasonably accurate contacts in and of themselves do not guarantee that the native topology will be found; but in all cases of accurate contacts, if there are a sufficient number of such contacts, then rather low RMSD structures are found in the pool; see Table VI. Thus, this is a reasonably effective method of predicting acceptably accurate tertiary contacts.

TABLE III
Predicted Contact Accuracy from Threading for 28 Proteins Used in an *Ab Initio* Folding Test[a]

Name of Protein	Number of Contacts Predicted	δ = 0	δ = 1	δ = 2[b]	δ = 3
1stfI	25	0.28	0.48	0.8*	0.88
1poh_	37	0.3	0.54	0.7*	0.7
1pou_	30	0.33	0.47	0.73*	0.9
1ife_	56	0.18	0.39	0.54	0.79
2azaA	47	0.38	0.53	0.79*	0.85
256bA	1	0	0	1*	1
1tlk_	53	0.81	0.94	1*	1
2pcy_	45	0.4	0.51	0.91*	0.91
1tfi_	52	0.19	0.35	0.60	0.79
2sarA	29	0.21	0.55	0.76	0.86
5fd1_	23	0	0.17	0.30	0.52
1cewI	7	0.57	0.86	0.86	0.86
1ctf_	46	0.11	0.3	0.50	0.7
1mba_	12	0.58	0.67	0.67	0.75
1shaA	41	0.34	0.66	0.85*	0.88
1thx_	53	0.23	0.55	0.72*	0.83
1shg_	42	0.19	0.57	0.76	0.86
1ubi_	23	0.61	0.65	0.78	0.83
6pti_	54	0.26	0.56	0.61	0.8
1cis_	19	0.21	0.58	0.95	0.95
1fas_	22	0.27	0.59	0.77	0.86
1ftz_	18	0.5	0.72	0.78*	0.89
1c5a_	20	0.1	0.3	0.4*	0.5
1fc2C	18	0.44	0.78	0.83*	1
1gpt_	19	0.37	0.53	0.79	0.89
lhmdA	33	0.18	0.36	0.52*	0.73
lixa_	14	0.43	0.64	0.79*	0.86
llea_	23	0.3	0.52	0.74*	0.96
Average		0.31	0.53	0.73	0.83

[a]δ = m is the number of contacts predicted within ±m residues of a correctly predicted contact. Correlated mutation analysis is from the CASP3 predictions of Ortiz et. al. [133].
[b]An asterisk indicates that this protein is foldable by *ab initio* (see Section VI).

V. GENERALIZED COMPARATIVE MODELING

Quality sequence-to-structure alignments generated by the threading procedure depend on the level of sequence identity of the target and the template proteins. In the cases of high sequence similarity, the protein folds are very similar, and classical methods of comparative modeling [194,195] led to good-quality models, frequently to models of similar quality to those obtained from the refinement of the X-ray data or good NMR data. When the sequence similarity becomes low or nondetectable by sequence comparison methods, the template proteins could be weakly homologous or just analogous—that is having similar folds without any obvious evolutionary relations. As a consequence, the resulting alignments are usually incomplete, with a substantial number of gaps and insertions. A fraction of residues of the probe protein, which is sometimes substantial, are not aligned to the template. Moreover, in the aligned parts of the structure, the true structure of the probe protein may differ in many important details from the structure resulting from the alignment to the template. Also, an optimal structural alignment of the two structures could be quite far from the threading-based alignment. Due to low sequence similarity, the threading alignment might not be the optimal one.

Is it possible to build a good-quality model based on poor alignments? Usually, it is not possible by means of contemporary procedures for comparative modeling. When the template structure differs substantially from the probe structure, the resulting models are typically much closer to the template structure than to the true structure of the probe protein [196]. The models do not move (in conformational space) in the direction of the probe structure, but instead wander around the template structure. Moreover, in the cases of large gaps in the alignment, the filled-in pieces of structure are sometimes completely nonphysical (non-protein-like).

A recently proposed method is described in the next sections that attempt to address this problem. The idea is to perform a kind of *ab initio* folding in the vicinity of the template structure, with the model force field controlling details of the folding. The template is used only to reduce the searchable portion of conformational space and loosely defines the general topology of the probe protein fold. The lattice model employed in these procedures has a limited resolution and accuracy. Consequently, the obtained models, in general, cannot achieve the accuracy of the experimental structures. As a result, it is rather pointless to apply the proposed methodology to those cases when the alignments are very good and complete. In such cases, the obtained structures would be slightly worse than structures built by classical comparative modeling tools. Such situations could be easily detected. In the remaining cases of low homology (or just analogy of the folds), the method is robust in the sense that it does not do any "harm" to the initial threading-based models and, for a substantial fraction of cases, leads to a qualitative improvement of the models. The resulting structures move toward the true probe structure. Because this approach bears some similarity to the comparative modeling, we call this method of homology/analogy-based structure prediction generalized comparative modeling (GeneComp, GC). The applied methodology is essentially the same for the template-restrained folding as for purely *ab initio* folding; the crossover is smooth, and there is no sharp boundary between threading-based and *ab initio* approaches.

A. Description of the Method

The method of generalized comparative modeling consists of several steps, which sequentially transform the threading alignment into a full-atom model of the probe protein. They are the following:

1. Build the threading alignment by a method described in the previous sections.
2. Construct the starting lattice model using the partial template from the threading as a structural scaffold.
3. Fold/optimize the lattice model using the threading alignment as a loosely defined structural template.
4. Cluster the lattice folding results [197] and/or calculate a mean structure by means of distance geometry (DG).
5. Refine the averaged model by Monte Carlo simulated annealing of an intermediate resolution off-lattice continuous model.
6. Reconstruct atomic details.

B. The Lattice Model and Its Force Field

Before describing the particular steps of the comparative modeling methodology, we outline the lattice model employed in all coarse-grained simulations (restrained or *ab initio*). Due to assumed reduced representation, we have named this protein model the side-chain-only (SICHO) model [139,198]. Technical details of the model design and its force field could be found elsewhere [199]. Here, an outline is provided for the reader's convenience. Most of the reduced models of proteins assume a more or less explicitly reduced (all-complete) representation of the main-chain backbone [200]. Frequently the alpha-carbon trace is used to represent the main-chain conformations, and the side chains are neglected or represented on various levels of simplification. When designing the present model, two partially contradictory goals were taken into consideration. First, for computational simplicity, there should just be a single degree of conformational freedom per residue. Second, the model should enable straightforward implementation of as accurate and selective a force field as possible. Thus, we assumed a single center of interactions that corresponds to the center of mass of the side group and the alpha carbon atoms.

This side-chain representation has several advantages over the alpha-carbon reduced representation. It is known that the sequence-specific interactions in proteins are due to different character of the side chains. The interactions of the main chain are rather generic. Then, having the coordinates of the side chains, it is very easy to reconstruct the main chain-coordinates [200]. In contrast, the reconstruction of the side-chain positions from the positions of the main chain is not trivial [201] and requires extensive optimization. Additionally, the side chains are bigger and their size varies between amino acids. Thus, this side-chain representation provides for better and more protein-like packing, with a well-defined first coordination shell.

The model chain is restricted to an underlying simple cubic lattice with the lattice spacing 1.45 Å. The set of possible virtual bonds between consecutive side chains is defined by a set of 646 lattice vectors. The shortest are of the vector type |±3,0,0| and |±2,±2,±1| while the longest are of the type |±5,±2,±1|, expressed in lattice units. The distribution of the length of the chain bond covers the majority (except for the wings) of the distribution seen in proteins. The main excluded volume is simulated by a cluster of the 19 closest (to the center of the model side chain) points on the underlying cubic lattice. This hard core of the chain is supplemented by soft-core repulsion spheres for the larger amino acids. The size of these spheres is adjusted in such a way that the folded model chains mimic average packing density of globular proteins.

The force field of the model consists of three types of potentials. First are the generic contributions that are independent of sequence and enforce the protein-like chain stiffness and internal packing. Potentials of the second type are amino acid-dependent and are used to reproduce the short-range interactions describing secondary structure propensities and orientation-dependent pair interactions. The potentials of the third type (short-range potentials identical in form to that described above and pairwise potentials [202]) are protein-dependent. Their derivation involves multiple sequence alignments of the sequence of interest, and the strength of interactions depends on the sequence similarity of protein fragments.

C. Construction of the Starting Lattice Chain

The threading alignment was used as a template to construct the initial lattice models. First, the aligned parts of the probe sequence were fitted to the template, and pieces of the lattice chain were built by taking into consideration the excluded volume of the model chain and the necessity of "stretching" the chain between the gaps in the template. Then, starting from the shortest loop, the loops and nonaligned chain ends were randomly inserted, again taking into account the excluded volume. The proper geometry of the model chain (avoiding nonphysical distances between side groups close along the chain) was preserved during the chain-building procedure. For good alignments, this procedure produces good models that need very little refinement. For extremely bad alignments, it may fail; in these (very rare) cases a less restrictive algorithm that allows for a larger deviation from the template could be used.

D. Restrained Lattice Folding: Optimization of the Initial Model

As discussed in Section II. D, the replica exchange Monte Carlo method appears to be an efficient tool for searching the conformational space of reduced protein

models. This technique was therefore used for the restrained folding (or refinement) of the probe proteins using the threading alignments as loosely defined structural templates. In the beginning of the procedure, a number of copies of the initial model are created and placed at various temperatures, according to the REMC scheme. Two subsequent runs were performed. In the first run, the range of temperatures is wider and shifted toward higher temperatures to allow for the fast equilibration of all replicas. In the subsequent longer run, the temperature range was smaller so that approximately half of the replicas run below the folding temperature and half above. About 20 replicas were usually simulated. This number of copies guarantees very fast and efficient swapping of conformations between the various temperature levels (the temperature increment between replicas has been assumed to be temperature-independent—a linear temperature set). A somewhat larger number of replicas may be required for fast convergence of larger proteins—250 residues or more. The conformations seen at the lowest temperature of the REMC scheme rapidly find the global energy minimum.

Three types of restraints are used to keep the sampling process in a broad conformational neighborhood of the template conformation.

The first is the most straightforward. The aligned portion of the template structure is placed at the center of the Monte Carlo working box. Then, at the beginning of the simulation, the starting chains are superimposed on the template. During the simulations, there are weak and somewhat ambiguous attractions (linear with distance) between aligned (according to the threading results) residues of the template and the moving probe chain. Thus during the simulation, the initial alignments have the chance to be corrected or even overridden by the model force field.

The set of tertiary contacts predicted by threading comprise the second set of restraints. Because only about one-third are correct and a much larger fraction are "almost" correct (i.e., they are shifted by ±1 or ±2 residues), the energy of attraction between the two residues of the probe predicted to be in contact grows linearly with the closest distance between the ±2 segments of the model chain. For very good alignments, the predicted contacts are, to a large extent, consistent with the template structure, and this set of restraints is essentially redundant to the restraints of the first type. For poorer alignments, a number of other locally similar proteins may contribute to the contact prediction. Consequently, the predicted contacts may significantly modify the resulting structures of the probe with respect to the template; that is, an averaged effect of other weak "templates" is introduced.

The third set of restraints contains the probe distances predicted from the fragment threading procedure. The distance restraints are limited to the pairs of residues that are no farther away than the length of the largest secondary structure element in the protein, which is equivalent to the estimated diameter (from the number of residues) of the probe protein.

E. Building the Average Models

For each probe protein, several independent simulations (10–20) were executed. From each simulation in the second pass, 200 conformations were stored in a constant interval of simulation time. The collected structures were averaged using a two-step distance geometry (DG), procedure. After the first pass, those structures far away from the average were rejected, and the final DG conformation was constructed from the remaining set of structures. Interestingly, DG averaging always led to a lower RMSD from the native than the average RMSD for the original set of conformations from the lattice simulations. Sometimes the structures from DG were close to the best structures seen in the folding simulations. Alternatively, our recently developed clustering procedure [197] could be used to identify clusters of the lowest energy conformations. The centroid of this cluster can then be treated as an averaged model. In the case of generalized comparative modeling, the two approaches are essentially equivalent. However, for *ab initio* folding, the clustering procedure is more powerful in identifying the most plausible fold from the sometimes-diverse results of *ab initio* lattice-folding simulations.

F. Reconstruction of Detailed Atomic Models

A very fast procedure was designed for reconstruction of the atomic details from the known positions of the alpha carbons and the side chains. The only constraints are the positions of the side-chain centers of mass. The initial local alpha-carbon trace geometry that is approximately reconstructed from the SICHO center-of-mass positions is not perfect. Therefore, the positions of alpha carbons are optimized in the first step. This is done by a gradient-optimization procedure using a very simple force field to improve the local geometry. At the next stage, positions of backbone atoms are reconstructed according to the local Cα trace conformation. In this step, the vector normal to the plane defined by three consecutive alpha carbons is calculated. This vector is almost parallel to a peptide bond plane. Thus, the remaining atoms of the peptide bond can be positioned quite accurately. Next, positions of side chain atoms are rebuilt. The conformations of the side chains are chosen from a representative database of rotamers. For rigid amino acids (e.g., phenylalanine), there is a single conformation in the database. There are up to 20 conformations for large, flexible side chains (e.g., lysine). The conformation of the rotamer depends on (a) the distance between the Cα atom and the center of mass of the side chain and (b) local chain conformation (i.e., Cα–Cα–Cα angle). Next, as a final stage of the reconstruction procedure, the side chains are rotated around a virtual Cα–center-of-mass bond—to avoid excluded volume conflicts. This procedure produces reasonable structures; however, the packing of side chains after all-atom reconstruction is not optimized. This can be done by one of the standard procedures of molecular mechanics. For the data reported in this work, this step was omitted.

G. Summary of Results on Fischer Database and Comparison with an Earlier Version of Generalized Comparative Modeling

Fischer's database of protein sequences and structures [34] is a standard benchmark set for validation of threading approaches. As mentioned previously, PROSPECTOR recognizes a majority of the related sequences correctly. Here, we would like to test our generalized comparative modeling approach on the same test set. Probably, Fischer's database [34] provides a very good test for the method. It contains closely related pairs of proteins (typical of homology modeling cases), pairs of weakly related proteins, and some pairs of very weakly similar ones. As suggested above, one may expect that for very closely homologous pairs of proteins, our method is not recommended. Indeed, the geometrical fidelity of the lattice model is in the range of 1 Å, and the model accuracy (due to deficiencies of the force field and to other factors associated with the reduced character of the model) is probably significantly lower and could be estimated to be about 2–3 Å. Also, for very weakly analogous proteins, where the template structure is far away from the probe structure and when the alignment is sparse or when alignment covers only a small fraction of the probe sequence, the method applied here will not provide good models: The restraints from the template prohibit the requisite large-scale rearrangements of the modeled structure. In most intermediate cases, one may expect a qualitative improvement of the model with respect to the quality of the initial threading-based models.

The above expectations are based on an earlier version of the generalized homology modeling with lattice folding in the neighborhood of the template structure [199]. The test results of the earlier approach are summarized in Table IV where an automated modeling by Modeller [203] (using the threading templates as starting points) is compared with lattice modeling refined by Modeller. While the number of cases given in this table is small, one may conclude that in a fraction of cases the improvement of the threading models is of a qualitative nature. Also, as expected, already-good models (see the example of 1aba_) do not improve. The threading procedure [181] used to generate the initial alignments for these 12 pairs produced worse alignments on average than the PROSPECTOR threading algorithm employed for the more massive test involving Fischer's database. To make the comparison more complete, for the few pairs that were not properly detected by PROSPECTOR, the match (and resulting alignments) was enforced, that is, the highest-scoring structural match was not taken as a template, but rather the correct structural template was used. The results for the proteins from Fischer's database are compiled in Table V.

TABLE IV

α-Carbon RMSD from Native for Models Built from the Initial Threading Alignments and Refined by Lattice Simulations[a]

Probe/Template Proteins	Threading + Modeller	SICHO + Modeller
1aba_/1ego_	4.43	4.86
1bbhA/2ccy_	6.77	6.82
1cewl/1molA	14.96	14.38
1hom_/1lfb_	7.82	3.70
1stfl/1molA	6.40	5.95
1tlk_/2rhe_	7.23	4.17
256bA/1bbh_	6.09	4.36
2azaA/1paz_	21.95	10.77
2pcy_/2azaA	6.56	4.41
2sarA/9rnt_	10.28	7.83
3cd4_/2rhe_	6.74	6.39
5fd1_/2fxd_	25.67	12.40

[a]The first column gives the PDB codes of the probe and template proteins detected by the threading algorithm. The second column gives the results of automated comparative modeling using the threading alignments as a template definition. The RMSD is given for the alpha-carbon trace. The right column contains the results of SICHO modeling followed by a refinement using the Modeller program. In the refinement stage the lattice models were used as a "template" for Modeller. Original alignments are the same for both approaches compared in the table.

Similar to the earlier version [199] of the comparative homology modeling, there are essentially three possibilities. First, when the threading model is very good the lattice modeling does not improve the overall quality of the molecular model; however, "no harm" to the quality of the model by application of the entire methodology could be assumed. Then, there are cases of topologically correct templates with moderate overall distance from the true probe structure. Here, in most cases a qualitative improvement of the model quality could be observed. Finally, for very bad initial models the final models are still not satisfactory; the accuracy is too low to be sure that the overall fold has been properly recovered. Some of these models can even contain topological errors.

A number of very interesting observations can be extracted from analysis of the data compiled in Table V. The first is that the lowest energy criterion for selection of the final model is not the best one. On the contrary, the distance geometry averaging or clustering procedures almost always provide models of better accuracy. The two methods (DG and clustering) lead to essentially the same (on average) quality of molecular models and are quite consistent. At the same time, it should be pointed out that the structure selection is not perfect. Usually the structures generated by clustering or DG are worse than the best structures observed in simulations. Definitely, better methods of selection (for example, based on all-atom structures) of the best structures from the lattice folding trajectories need to be developed.

TABLE V

Compliation of Results of Generalized Comparative Modeling on Proteins from the Fischer Database[a]

Target	Template	Alignment Coverage	Aligned Part	Best RMSD	Lowest Energy	DG	First Cluster
1aaj_	1paz_	82.86	6.74	6.15	9.26	9.37	9.00
1aba_	1ego_	90.81	6.52	3.55	5.90	4.75	3.95
1aep_	256bA	64.05	18.36	18.31	18.36	21.45	22.38
1arb_	4ptp_	80.99	16.32	15.78	17.47	17.46	17.69
1atnA	1atr_	75.27	12.42	12.00	13.25	13.16	13.04
1bbhA	2ccyA	93.89	2.74	2.71	3.65	3.07	2.99
1bbt1	2plv1	93.59	12.55	9.57	10.81	10.70	10.80
1bgeB	1gmfA	66.67	7.89	4.93	6.27	5.45	5.71
1c2rA	1ycc_	85.35	4.35	4.31	5.75	5.34	5.30
1cauB	1cauA	89.63	5.18	4.04	5.69	5.45	5.41
1cewI	1molA	70.37	4.85	4.10	8.00	7.79	7.83
1chrA	2mnr_	92.97	3.50	3.77	5.35	4.90	4.78
1cid_	2rhe_	55.93	19.76	14.05	18.88	18.44	16.97
1cpcL	1colA	81.40	15.71	12.30	13.43	13.58	13.17
1crl_	1ede_	47.75	20.01	21.35	24.21	24.09	24.93
1dsbA	2trxA	51.65	12.46	11.58	15.94	16.47	15.30
1dxtB	1hbg_	92.52	2.74	2.91	3.54	3.01	3.08
1eaf_	4cla_	78.13	13.25	9.27	10.09	10.32	10.10
1fc1A	2fb4H	96.62	12.99	2.63	3.21	13.12	2.74
1fxiA	1ubq_	61.46	10.94	8.53	10.28	10.18	10.14
1gal_	3cox_	74.01	15.03	14.03	17.74	17.80	17.38
1gky_	3adk_	85.48	6.68	6.13	8.75	6.36	8.87
1gp1A	2trxA	54.89	11.48	9.08	14.75	13.74	15.06
1hip_	2hipA	80.00	3.55	3.92	4.86	4.26	4.13
1hom_	1lfb_	97.73	1.62	1.50	2.30	1.57	1.70
1hrhA	1rnh_	91.30	7.15	4.90	5.50	5.07	5.07
1isuA	2hipA	95.16	6.06	3.20	4.35	5.07	4.08
1lgaA	2cyp_	77.60	12.45	12.44	17.14	15.59	16.53
1ltsD	1bovA	59.00	9.99	8.11	12.16	10.21	9.47
1mdc_	1ifc_	96.97	2.62	2.55	3.12	2.66	2.65
1mioC	1minB	88.38	14.48	14.05	15.19	14.71	14.94
1mup_	1rbp_	93.63	5.56	4.14	4.89	4.38	4.51
1npx_	3grs_	92.17	14.56	13.61	14.15	14.12	14.09
1onc_	7rsa_	98.08	3.81	3.08	3.53	3.51	3.29
1osa_	4cpv_	70.27	16.84	16.56	18.02	17.90	17.81
1pfc_	3hlaB	89.22	3.84	3.81	4.69	4.28	4.46
1rcb_	1gmfA	71.32	6.28	3.91	5.51	6.09	4.25
1sacA	1ayh_	76.47	18.13	16.89	18.52	18.81	18.93
1stfI	1molA	69.47	8.46	4.97	7.38	7.07	8.11
1tahA	1tca_	56.92	19.00	18.90	21.60	21.51	20.96
1ten_	3hhrB	93.33	5.60	3.14	3.98	3.62	3.45
1tie_	4fgf_	66.87	7.88	7.88	8.80	8.60	8.94
1tlk_	2rhe_	95.83	4.61	2.35	3.49	3.42	3.03
2afnA	1aozA	95.83	25.27	22.60	23.68	25.05	23.50

TABLE V (*Continued*)

Target	Template	Alignment Coverage	Aligned Part	Best RMSD	Lowest Energy	DG	First Cluster
2ak3A	1gky_	78.26	15.63	14.65	15.51	15.46	15.27
2azaA	1paz_	62.79	7.60	6.33	8.40	7.87	7.30
2cmd_	6ldh_	95.83	5.02	4.22	4.74	4.44	4.49
2fbjL	8fabB	94.37	10.30	7.04	7.72	8.78	8.37
2gbp_	2liv_	80.94	10.72	9.50	10.66	10.07	10.35
2hhmA	1fbpA	71.69	15.26	15.99	18.30	17.57	17.83
2hpdA	2cpp_	85.33	6.44	5.41	6.75	5.83	5.81
2mnr_	4enl_	95.52	14.92	13.55	14.07	14.28	14.27
2mtaC	1ycc_	65.31	14.35	14.04	16.01	16.49	16.51
2omf_	2por_	82.06	23.61	21.82	23.51	23.45	24.17
2pia_	1fnr_	79.44	15.72	15.64	17.29	16.77	18.24
2pna_	1shaA	46.55	10.69	7.27	11.31	8.92	10.89
2sarA	9rnt_	91.67	6.36	4.88	6.11	5.76	5.84
2sas_	2scpA	86.49	6.45	5.51	6.42	6.11	5.95
2sga_	4ptp_	98.82	17.74	9.78	11.87	10.49	11.94
2sim_	1nsbA	66.14	14.34	16.52	19.79	18.57	17.47
2snv_	4ptp_	84.11	14.28	12.78	14.07	13.84	13.31
3cd4_	2rhe_	92.78	7.02	5.98	7.40	7.15	7.05
3chy_	4fxn_	86.72	6.07	3.58	4.91	4.36	4.59
3hlaB	2rhe_	83.15	10.30	4.72	9.76	8.63	8.62
3rubL	6xia_	74.13	20.91	22.26	24.19	24.15	23.71
4sbvA	2tbvA	97.49	18.68	17.73	18.47	18.53	18.97
5fd1_	2fxb_	55.66	10.95	10.70	12.13	11.99	11.61
8i1b_	4fgf_	73.97	11.31	10.77	12.58	12.88	12.65

[a]The first two columns contain the PDB codes of the target and template proteins, respectively. The percentage of a target sequence aligned to a template is given in column 3. The fourth column provides RMSD (all values for alpha-carbon traces) for the aligned part of the template from "true" structure of the target—a measure of the alignment quality. The fifth column gives the best RMSD for the model chains observed in a set of sparely written trajectories (a few hundred photographs). The sixth column gives the RMSD for the lowest energy (according to the SICHO force field) conformation observed in the trajectories. The RMSD values in the two last columns correspond to the average structures obtained via distance geometry and clustering algorithm. The two methods of averaging are almost equivalent, with slightly better performance of the DG approach. In number of cases, the final models for the entire structure are better (as measured by RMSD from the crystallographic structure) than the initial threading models—that is the aligned part.

H. Comparison to Modeller

Recently, several tools were developed for the fast building of all-atom models of proteins by various means of comparative modeling. Probably, the most efficient is Modeller, developed by Sali and Blundel [195]. Modeller allows for the high-throughput modeling of protein structures on a genomic scale. The method proposed here is more complex and more computationally demanding; however, it is still feasible in large-scale applications. The key question is, Are the results worth the increased computational cost? To answer this question, we compared various models for the Fischer database proteins [34] in Table VI, where the results of generalized comparative modeling described in this contribution are compared with models generated by Modeller. Both procedures started from exactly the same templates and the same alignments generated by PROSPECTOR. If we consider all models, then GeneComp performs better than Modeller in 53 cases, worse in 13, and the same in two cases. If only templates whose RMSD is less than 10 Å are considered, then GeneComp performs better in 29 cases, Modeller performs better in five cases, and they perform the same in one case. However, in the latter, the two structures differ by a small amount. In many cases of very good (or good) templates, the two methods generate models of similar quality. The situation changes when the homology becomes weaker and when, consequently, the threading models become more distant from the probe structure. In these cases, the models generated by GeneComp are almost always of noticeably better accuracy. We can most likely ignore the cases when both methods lead to very bad models. It is safe to say that there is usually no difference between models 12 and 14 Å from the true probe structure. The utility of such models for structural genomics is at least problematic (of course, it depends somewhat on protein size—a very large protein may still be of a correct overall topology with this high RMSD). However, there is quite a difference between a model that is 4 Å from the true structure and a 6 Å model (or even more between a 6 Å model and 10 Å model). As can easily be seen from the data compiled in Table VI, in the range of 4–8 Å, the GeneComp models are in most cases significantly more accurate than the models generated by Modeller. The typical difference is 1–2 Å; however, in a few cases it is as much as 4–5 Å. Interestingly, the models generated by GeneComp frequently have a lower RMSD for the entire structure than the RMSD of the original aligned fragments. These are the cases when a qualitative improvement with respect to simple comparative modeling was observed. The lattice simulations improve entire structures. Thus, on average the proposed method leads to qualitatively better molecular models with pronounced consequences for structure-based protein function prediction and other aspects of proteomics.

VI. *AB INITIO* FOLDING

A. Description of the Method

The method for *ab initio* folding of small globular proteins employs the same modeling tools as in generalized comparative modeling. There are, however, some differences. Of course, now there is no template to restrict the

TABLE VI
Comparison of Generalized Comparative Modeling with Automated Modeling via Modeller[a]

Target	GeneComp + DG	Modeller	GeneComp + DG + Modeller
1aaj_	9.37	10.13	9.30
1aba_	4.75	6.66	4.73
1aep_	21.45	21.56	21.32
1arb_	17.46	18.56	17.35
1atnA	13.16	15.61	13.15
1bbhA	3.07	3.02	3.03
1bbt1	10.70	10.21	10.68
1bgeB	5.45	10.34	5.42
1c2rA	5.34	5.84	5.30
1cauB	5.45	5.93	5.93
1cewI	7.79	8.47	7.76
1chrA	4.90	4.57	4.91
1cid_	18.44	20.19	18.44
1cpcL	13.58	15.62	13.52
1crl_	24.09	25.89	23.98
1dsbA	16.47	16.37	16.45
1dxtB	3.01	3.05	3.00
1eaf_	10.32	10.82	10.18
1fc1A	13.12	15.02	12.48
1fxiA	10.18	11.27	10.11
1gal_	17.80	18.86	17.66
1gky_	6.36	11.82	6.45
1gp1A	13.74	15.22	13.66
1hip_	4.26	4.06	4.09
1hom_	1.57	1.73	1.57
1hrhA	5.07	6.95	5.05
1isuA	5.07	5.84	5.20
1lgaA	15.59	14.72	15.68
1ltsD	10.21	10.88	10.22
1mdc_	2.66	2.66	2.71
1mioC	14.71	16.78	14.68
1mup_	4.38	4.93	4.40
1npx_	14.12	14.48	14.05
1onc_	3.51	5.14	3.50
1osa_	17.90	16.89	17.91
1pfc_	4.28	4.39	4.49

[a] The same alignments (see Table V) were used as starting templates for GeneComp (RMSD for the DG averaged models) and Modeller. The last column provides RMSD for the models generated by Modeller starting from the complete models obtained by GeneComp. In almost all cases the models generated by GeneComp are more accurate than the models generated by Modeller, and in 15–20 cases the improvement is of a qualitative nature (see the text for explanation). Refinement of the GeneComp models by Modeller (compare columns 2 and 4) leads to marginal changes of the molecular models, indicating the consistency of the GeneComp models, with local atomic details of the PDB structures.

conformational search. The generic and protein-independent components of the force field for the lattice models are the same, and the protein-specific potentials have a similar form [202]. The difference is that in *ab initio* folding they are less specific. For the test purposes, all homologous (and analogous) proteins have been excised from the structural database used to derive the potentials. As a result, the number and accuracy of the predicted contacts are lower, as is the accuracy of the short-range terms. As before, a conservative prediction of the regular elements of secondary structure was used to bias the short-range interactions. Thus the requirements for the folding simulations are much higher. A much larger number of independent simulations were executed to check the reproducibility of the results and to provide a representative sample for the clustering procedure and final fold selection.

The selection of the initial conformations for the REMC simulations requires some comment. In principle, random expanded conformations could be used. However, this slows down the convergence of the process. For this reason, a different strategy was adopted. Having a prediction of secondary structure, gapless threading of structures of comparable size is performed using the matching fractions of the predicted secondary structure to the actual secondary structure of the templates as a scoring function. Of course, all homologous and analogous proteins were removed from the pool. Fifty lattice chains were built using the 50 best scoring structures as templates. While these starting structures are different from the probe fold, they may have the proper element(s) of secondary structure that may serve as a fast nucleation site for the folding process. In the preliminary simulation runs, 50 replicas were used. The second iterations used the top 20 (20 lowest-energy replicas) as the input pool. The simulation results from the last iteration of the lattice-folding algorithm were subject to a clustering procedure [197] that was also used to make the final fold selection.

B. Results of *Ab Initio* Folding on 28 Test Proteins

Sequences of 28 globular proteins were selected as the test set for the *ab initio* folding protocol. The set is representative of single-domain small proteins. It contains alpha proteins with α/β-, $\alpha + \beta$-, and β-type folds. In about 50% of the cases, low-resolution folds of correct topology were obtained as one of a number of clusters. The results are compiled in Table VII that also contains the RMSD for the best structures observed during simulations at the lowest temperature replica of the system as well as the RMSD of all structures that cluster [197]. It is clear that simulations generate a small subset of very good structures for the majority (22 of 28) of the tested proteins. Unfortunately, the fold selection procedure rarely selects structures close to the very best ones. The discrepancy is more drastic than in the case of template-restricted folding. It could be proven rigorously that to obtain a 3 Å structure by random in a set of trajectories

TABLE VII
Summary of *Ab Initio* Folding Results

Protein Name[a]	Best RMSD	Lowest-Energy RMSD	RMSD of Centroid of Each Cluster
1c5a_	4.86	10.87	11.20 11.63 **5.70** 8.75
1cewI	6.71	10.08	8.77 13.84 15.29 12.00 11.66
1cis_	4.98	11.52	10.41 10.34 9.36 9.67 10.43 6.81 7.25
1ctf_	7.10	11.06	10.72 11.40 11.54
1fas_	5.30	8.55	9.30 7.47 11.68 10.15 11.89 **6.36** 12.87
1fc2C	2.91	7.34	7.21 7.61 **3.35**
1ftz_	2.65	8.79	8.78 6.52 **3.05** 7.11 6.50 8.18
1gpt_	4.92	7.45	7.58 8.66 9.70 9.59
1hmdA	5.02	10.57	10.36 12.95 14.20 12.52 **5.51**
1ife_	6.53	9.23	11.57 9.24 13.64 11.71 12.12 11.41
1ixa_	4.02	6.62	**6.36** 6.92 9.28 10.65 10.53
1lea_	3.23	11.85	10.93 9.95 8.32 8.44 **5.82**
1mba_	9.61	12.72	12.63 15.28 12.01 15.44 13.51
1poh_	2.90	12.63	12.76 11.91 **3.87**
1pou_	2.70	4.98	**3.95** 9.88 9.93 10.93 11.61
1shaA	3.94	13.07	13.82 12.08 12.75 9.00 10.49 6.00
1shg_	4.40	9.00	8.99 9.06
1stfI	5.47	10.19	8.06 12.86 11.17 13.68 11.99 16.74
1tfi_	7.62	9.48	10.15 8.88 10.56 10.20
1thx_	2.97	12.72	12.83 11.27 **3.89** 13.04 14.40
1tlk_	3.13	7.38	11.02 **6.35**
1ubi_	3.05	10.98	10.71 10.51 11.57 12.07 8.13 10.54
256bA	3.09	3.73	**3.52** 8.38 14.88 10.01 14.91 12.13
2azaA	3.83	7.20	**5.75 12.86 13.01 14.00 13.30 13.30**
2pcy_	3.72	7.75	**5.56** 7.12 11.39 13.46 13.19
2sarA	8.45	13.11	10.71 11.92 12.18 12.71 14.10 13.93 14.10 13.79
5fd1_	8.67	12.53	12.20 10.84 12.48 10.94 14.35 14.26
6pti_	5.36	7.36	6.68 10.81 10.99 10.14 9.14

[a] Bold indicates that this protein is foldable; that is, one of the clusters has an average RMSD from native less than 6.5 Å.

containing a few thousand photographs is practically impossible. Thus, the model force field and the sampling scheme do a reasonably good job in sampling protein-like regions of conformational space, including the neighborhood of the native state. At the same time, the force field lacks a sufficient discriminatory ability to select the closest-to-native fold generated from a large number of competing protein-like structures. These competing structures have elements of native topology with misfolded fragments of structure; sometimes they are mirror images of native-like folds.

Overall, though, if one defines a successful simulation as one with a native topology whose backbone RMSD is less than 6.5 Å, then in 15/28 cases (i.e., about 54% of the cases) the simulations are successful. Again, a different, more efficient fold selection method needs to be developed; such efforts are currently underway. An alternative recently being explored is the method of inserting atomic detail and then scoring the structures using a recently developed distance-dependent potential of mean force [204]. If this is done, then 1stfI is not foldable, but 1fas_,1gpt_,1mba_ are foldable, giving a total of 17 (i.e., 61%) of the test set proteins successfully folded.

VII. COMPATIBILITY OF REDUCED AND ATOMIC MODELS

A. Reproducibility of Structural Details

Reduced models have a long history. Some reproduce just the overall fold of globular proteins, whereas other (more complex) models maintain some details of protein structure. The SICHO model, based on just a single center of interaction per residue, appears at first glance to be a drastic simplification. However, due to its flexibility, the model is more accurate than it may appear at first. First of all, the mesh size of the underlying cubic lattice is equal to 1.45 Å, which means that a simple fit of the lattice model to a detailed PDB [171] structure has an average accuracy of 0.7–0.8 Å with respect to the side-chain centers of mass. Due to the coarse-grained character of the potentials, correctly folded (say, by a pure *ab initio* approach) structures are of somewhat lower accuracy. Very small proteins or peptides could be folded to 1.5 Å to 2.0 Å from the native structure. The accuracy of larger proteins decreases due to an accumulation of errors across the structure. For 100-residue proteins, properly folded structures have an RMSD in the range of 3.5–6.5 Å from native. When looking for elements of secondary structure as helices and β-hairpins, the accuracy is of the same range as for very small proteins or slightly better and ranges between 1.0 and 2.0 Å. The above numbers are given for the side-chain centers of mass. Our model employs a very crude and simple reconstruction of the α-carbon coordinates as a simple combination (with the coefficients extracted from a statistical analysis of the structural database) of the positions of three consecutive side-chain centers. This estimation is contaminated by a small systematic error (there is no correction from deviation of the α-carbon from the plane defined by three corresponding side-chain united atoms) and by some statistical error related to errors in the side-chain positions. Compensating for this is a statistical reduction of the absolute error of Cαs because the main-chain units are "inside" the secondary structure elements defined by the side-chain centers of mass. Consequently, errors in the side-chain positions translate into a slightly smaller error in the positions of the α-carbons. As a result, the accuracy of the crude α-carbon trace is the same or slightly better than the accuracy of the explicit virtual chain of the side groups.

The level of local (and global) accuracy of the model is sufficient to allow for quite accurate reproductions of the most important structural details. First, the contact maps of the side chains extracted from the model are very similar to the contact maps calculated from the crystallographic structures, assuming a 4.5 Å cutoff for contacts between heavy atoms of the side chains (side groups are considered to be in contact when any pair of their heavy atoms are at a distance smaller than the above cutoff). The overlap with native for properly folded structures is 85–90%. There are some excess contacts in the lattice models, and some contacts are missed due to the spherical shape of the model side chains and the statistical character of the cutoff distances for the model residues. More interestingly, the model hydrogen bond network (properly calculated from the estimated coordinates of alpha carbons) of the main chain coincides with similar (85–90%) accuracy with the main-chain hydrogen bonds assigned by the DSSP procedure [205] to the corresponding native structures. Bifurcated hydrogen bonds (the weaker ones) are ignored in this comparison, because the model does not allow for H-bond bifurcation. As in real proteins, the model structures have very regular networks of hydrogen bonds. Helices, except for their ends, exhibit a regular pattern of two hydrogen bonds per residue. The same is observed for internal β-strands in β-sheets. The edge strands usually have a single model H-bond per residue. Sometimes, even patterns characteristic of β-bulges are reproduced with high fidelity. The model network of H-bonds is explicitly cooperative. This leads to protein-like cooperative folding. Interestingly, misfolded structures also look very protein-like unless they violate some "rules" of protein folding—for example, the handedness of the β–α–β connections [206].

The protein-like geometry of such a simple model is enforced by the proper design of the force field that has two distinct types of components: sequence-dependent (or even protein-specific), which drive folding toward a specific fold, and generic, which strongly bias the model chain toward the average protein-like local conformational stiffness. The force field also has packing preferences. This way a vast majority of the irrelevant portion of the conformational space of the high coordination lattice (containing 646 possible side-chain–side-chain virtual bonds) model is efficiently avoided during the sampling process.

B. Reconstruction of Atomic Details

The lattice SICHO model exhibits good compatibility with detailed all-atom models. Projection of the all-atom structures onto the lattice model is trivial, and the accuracy of the projection is about 0.8 Å RMSD for the side-chain centers of mass or for the coarse reconstruction of all the α-carbon positions. More interesting, and certainly more challenging, is the reconstruction of the atomic details from the lattice models. A couple of similar procedures have recently been developed for this purpose [200]. In one, the crude estimated coordinates of the α-carbons are refined using the distance restraints typical for proteins and simple potentials for optimization of the backbone geometry. In the next stage, the remaining atoms of the main chain are reconstructed using a library of backbone fragments. Finally, a library of side-chain rotamers is employed to build the side-group conformations that are the most consistent with the lattice model. The side-group geometry and packing can be optimized relatively easily because the gross overlaps are by definition excluded by placing the rotamers as close as possible to the lattice chain (which itself exhibits a reasonable approximation of the packing in a protein). When starting from the lattice fit to the crystallographic structure, this reconstruction process returns a full atom structure that differs on average by about 1 Å RMSD from the original one. Further minimization by the CHARMM force field [207] leads to a small improvement of the model. The same accuracy of all-atom reconstruction is expected for all conformations generated during the lattice simulations.

A somewhat different procedure that has an advantage of computational speed leads to structures that are about 1.5 Å from the original all-atom model. Thus, there is the possibility of multiscale simulations of protein systems. The computational speed of the SICHO model enables simulations that correspond to the time-scales characteristic of real protein folding. At specific interesting points of MC trajectory, one can perform all-atom reconstruction, followed by detailed MD simulations. Another possibility that is now being explored is to use the all-atom models (derived from lattice structures) as a means of selecting the "best," possibly closest to native, structures generated in lattice folding simulations by the SICHO model.

C. Feasibility of Structural Refinement

As discussed in other parts of this chapter (see Sections VIII and IX), low-resolution models could be successfully employed in the functional annotation of new proteins and even for docking ligands. Of course, the more accurate the model, the wider its applications. The SICHO model is of limited resolution. Typical, well-folded structures have an RMSD that is 2 to 6.5 Å from native. Is it possible to improve such models using more a detailed representation and a more exact force field? Is it possible to include the solvent successfully in an explicit way at this stage? It appears that at least for moderately small proteins with a reasonable starting lattice structure, sometimes the models can be refined to a resolution close to that of experimental structures. Successful refinement of a small protein, CMTI, from a low-resolution MONSSTER folding algorithm [137] to a structure close to the experimental one was recently done by Simmerling et al. [208]. Earlier, for similar low-resolution lattice models, several structures of leucine zippers were also successfully refined to experimental resolution [124,125]. These studies were subsequently extended using ESMC to provide a treatment of the GCN4 leucine zipper folding thermodynamics as well as the prediction of the native state [209], and it was subsequently shown that the CHARMM force field, when supplemented by a generalized Born/surface area treatment, is highly correlated with the lattice-based force field [210]. These studies are extremely encouraging, although it is now unclear how soon the gap between low-resolution lattice folds and high-resolution all-atom structures for larger proteins will be closed.

VIII. FROM STRUCTURE TO BIOCHEMICAL FUNCTION

A. Does Knowledge of Protein Structure Alone Imply Protein Function?

Because proteins can have similar folds but different functions [211,212], determining the structure of a protein does not necessarily reveal its function. The most well-studied example is the $(\alpha/\beta)_8$ barrel enzymes, of which triose phosphate isomerase (TIM) is the archetypal representative. Members of this family have similar overall structures but different functions, including differing active sites, substrate specificities, and cofactor requirements [213,214]. An analysis of the 1997 SCOP database [211] shows that the five largest fold families are the ferredoxin-like, the (α/β) barrels, the knottins, the immunoglobulin-like, and the flavodoxin-like fold families with 22, 18, 13, 9, and 9 subfamilies, respectively. In fact, 57 of the SCOP fold families consist of multiple superfamilies [15]. These data only show the tip of the iceberg: Each superfamily is further composed of protein families, and each individual family can have radically different functions. For example, the ferredoxin-like superfamily contains families identified as Fe–S ferredoxins, ribosomal proteins, DNA-binding proteins, and phosphatases, among others. More recently, a much more detailed analysis of the SCOP database has been published [215], which finds broad function–structure correlation for some structural classes, but also finds a number of ubiquitous functions and structures that occur across a number of families. The article provides a useful analysis of the confidence with which structure and function can be correlated [215]. For a number of functional classes, knowledge of protein structure alone is insufficient information to assign the specific details of protein function.

B. Active Site Identification

It has been suggested that the active sites in proteins are better conserved than the overall fold [27]. If so, then one should be able to identify not only distant ancestors with the same global fold and same biochemical activity, but also proteins with similar functions but different global folds. Nussinov and coworkers empirically demonstrated that the active sites of eukaryotic serine proteases, subtilisins, and sulfhydryl proteases exhibit similar structural motifs [216]. Furthermore, in a recent modeling study of *S. cerevisiae* proteins, active sites were found to be more conserved than other regions [27]; this was also seen in the study of the catalytic triad of the α/β hydrolases [11]. Kasuya and Thornton [217] have created structural analogs of a number of Prosite sequence motifs and showed, for the 20 most frequent Prosite patterns, that the associated local structure is rather distinct [3]. These results provide clear evidence that enzyme active sites are structurally more highly conserved than other regions of a protein.

C. Identification of Active Sites in Experimental Structures

Several groups have identified functional sites in proteins with the goal of engineering or inserting functional sites into new locations, and success has been achieved for several metal-binding sites [218–226]. However, because highly accurate site descriptors of backbone and side-chain atoms were used, this fueled the idea that significant atomic detail is required if protein structure is to be used to identify protein function. Similarly, detailed side-chain active site descriptors of serine proteases and related proteins were employed to identify functional sites [227], while more automated methods for finding spatial motifs in protein structures have been developed [37,216,228–233].

Unfortunately, such methods require the exact placement of atoms within protein side chains and are inapplicable to the inexact, low-resolution predicted structures generated by the state-of-the-art *ab initio* folding and threading algorithms (see Sections IV–VI). These methods are required when the sequence identity of the sequence of interest to solved structures is too low to use comparative modeling. To address this need, Skolnick and Fetrow have recently developed "fuzzy," inexact descriptors of protein functional sites [8]. They are applicable to both high-resolution, experimental structures and low-resolution (backbone RMSD 4–6 Å from native) structures. These descriptors are α-carbon-based, "fuzzy functional forms" (FFFs). Initially, they created FFFs for the disulfide oxidoreductase [8,10] and α/β-hydrolase catalytic active sites [11] (an additional 198 have now been built, with comparable results [234]).

The disulfide oxidoreductase FFF was originally applied to screen 364 high-resolution structures from the Brookhaven protein database [235]. For the true positives, the proteins used to create the FFF have different structures and low sequence identity to those proteins used to build the FFF, but the active sites are quite similar [8]. Here, the FFF accurately identified all disulfide oxidoreductases [8]. In a larger dataset of 1501 proteins, the FFF again accurately identified all of the disulfide oxidoreductases, but it also selected another protein, 1fjm, a serine-threonine phosphatase. Initially this was a discouraging result, but subsequent examination of the sequence alignments combined with an analysis of the subfamily clustering strongly suggested that this putative active site might indeed be a site of redox regulation in the serine-threonine phosphatase-1 family [12]. If experimentally verified, this would highlight the advantages of using structural descriptors to analyze multiple functional sites in proteins. In particular, function prediction would not be restricted to the "primordial" function that characterizes the sequence family, but could also include additional functions gained during the course of evolution.

D. Requirements of Sequence–Structure–Function Prediction Methods

Any sequence–structure–function method that does function prediction by analogy relies on three key features. First, the function of the template protein must be known. Second, the active site residues must be identified and associated with the function of the protein. Third, a crystal structure of a protein that contains the active site must be solved so one can excise the active site for constructing the corresponding three-dimensional active site motif. Evolutionary approaches to function prediction often just require that the first criterion be satisfied, but for more distant homologs the second should be checked as well, because functions can be modified during evolution. The third requirement is unique to structure-based approaches to function prediction. Based on studies to date [8,10–12,14,15], identification of an enzyme's active site requires a model whose backbone RMSD from native near the active sites is about 4–6 Å for structures generated by *ab initio* folding. This predicted structure quality is due to the fact that the errors in the active site geometry found in the predicted structure tend to be systematic rather than random. However, threading does not suffer from this problem because, in the predicted structure, if the alignment does not include the active site residues, no functional prediction is made. If it does, the local geometry is the same as in the template's native structure. Threading can have alignment problems, but locally—at least in the vicinity of the active site—these can often be overcome if the threading score includes a sequence similarity component or if Generalized Comparative Modeling is done. Nevertheless, in practice, for both *ab initio* and threading models, the quality of the predicted structures is better in the core of the molecule than in the loops, so prediction of the function of a protein whose active site is in loops may be problematic. Currently, the method has only been applied to identify enzyme active sites. Recent work described in Section VIII suggests that at least in some situations, low-resolution structures can also be used to at least partially address the problem of substrate and ligand binding. But in general, techniques that will further refine inexact protein models will be necessary to extend the approach.

E. Use of Predicted Structures from *Ab Initio* Folding

As noted above, the recent CASP3 results suggest that for small proteins, current tertiary structure prediction schemes can often (but far from always) create inexact protein models of the global fold. Are these structures useful for identifying functional sites in proteins? To explore this issue, using the *ab initio* structure prediction program MONSSTER [191,193], the tertiary structure of the glutaredoxin, lego, was predicted whose backbone RMSD from the crystal structure was 5.7 Å. To determine if this inexact model could be used for function identification, the set of correctly folded structures and a set of 55 incorrectly folded structures were screened with the FFF for disulfide oxidoreductase activity [8,10]. The FFF uniquely identified the active site in the correctly folded structure but not in a library of incorrectly folded ones [15]. This is a proof-of-principle demonstration that inexact models produced by the *ab initio* prediction of structure from sequence can be used for the prediction of biochemical function.

F. Use of Threaded Structures to Predict Biochemical Function

In a very important paper, Lathrop demonstrated that use of functionally conserved residues could filter threading predictions to correctly identify globins even when the threading score was insignificant [30]. While suggestive, the key question was whether or not this result could be generalized on a genomic scale. Over the past few years, we have been exploring this issue in great detail [8,10–15], and, as discussed below, we demonstrate that the use of the sequence–structure–function paradigm, when appropriately employed, allows one to predict biochemical function with a much smaller false-positive rate than BLOCKS [236,237], the best competing sequence-based approach. Indeed, we have developed a very promising approach to the problem of genome-scale function annotation.

The methodology is as follows: We use PROSPECTOR1 [57] (although, any threading algorithm could, in principle, be used) to identify the set of 20 structures that are the best scoring matches between the probe sequence and the template structure (four scoring functions times five best scoring structures for each function). Then, each structure was searched for matches to the active site residues and geometry of the FFF. If a match to the FFF is found, then for those sequences for which homologous sequences are available, a sequence-conservation profile was constructed [11]. If the putative active site residues are not conserved in the sequence subfamily to which the protein belongs, that sequence is eliminated as having the predicted function; otherwise the sequence is predicted to have the function. Using this sequence–structure–function method, 99% of the proteins in the eight genomes that have known disulfide oxidoreductase activity were found [15]; 10% to 30% more correct functional predictions are made than in alternative sequence-based approaches [15]; similar results are seen for the α/β-hydrolases [11].

In Fig. 2, we show the distribution of scores (blue) for the *E. coli* genome [238] when any of the 11 disulfide oxidoreductases in our structural database was selected as being in the top five scoring structures using the "close" sequence plus secondary structure plus pair profile scoring function. Similarly, those proteins identified on application of the disulfide oxidoreductase FFF to

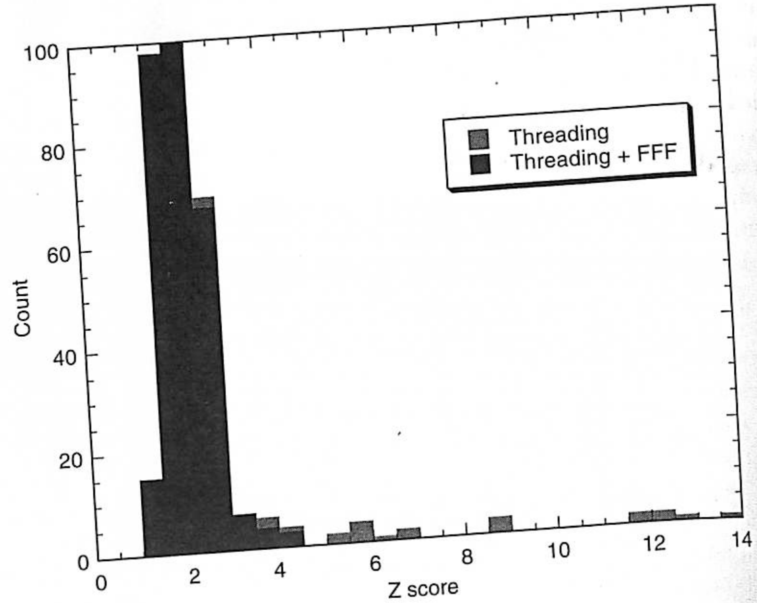


Figure 2. For the *E. coli* genome, the distribution of threading scores for the "close" sequence plus secondary structure/pair profile scoring function is shown in dark gray and those proteins identified by use of the disulfide oxidoreductase FFF are shown in light gray.

these threading models (all are known true positives) are indicated in red. Clearly, the use of the FFF allows one to extract proteins (e.g., those to the immediate right of the maximum) when their raw threading score would require one to also include a significant (in this case overwhelming) number of false positives. We note that full use of PROSPECTOR1-3 identifies all the known disulfide oxidoreductases in the *E. coli* and *M. genitalium* genomes. Note that, in general, structures whose *Z*-score is greater than 1 can be successfully searched for a match to a known active site.

Importantly, using structural information, the false-positive rate is much less than that found using sequence-based approaches. This conclusion arises from a detailed comparison of the FFF structural approach and the Blocks sequence-motif approach [15]. Here, the sequences in eight genomes, including *B. subtilis* [239], were analyzed for disulfide oxidoreductase function using the disulfide oxidoreductase FFF, the blocks thioredoxin block 00194 [236], and the blocks glutaredoxin block 00195 [236]. In Fig. 3 we plot the distribution of scores when the *B. subtilis* genome is threading through these two blocks. By way of example, if we assume that those sequences identified by both the FFF and Blocks [236] are "true positives," we find 13 such sequences in the *B. subtilis* genome. (Recognize that the experimental evidence validating all of these "true positives" is lacking; thus, they are more accurately termed "consensus positives.") To find these 13 "consensus positive" sequences, the FFF hits 7 false positives. In contrast, Blocks hits 23 false positives. It was previously suggested that the use of a functional requirement adds information to threading and reduces the number of false positives [30]. These data validate this claim on a genome-wide basis. Similarly, using active site descriptors as a filter, one can identify the true positives even when the threading score is barely significant (as in Fig. 2) and where selection of the structure based on the threading score alone

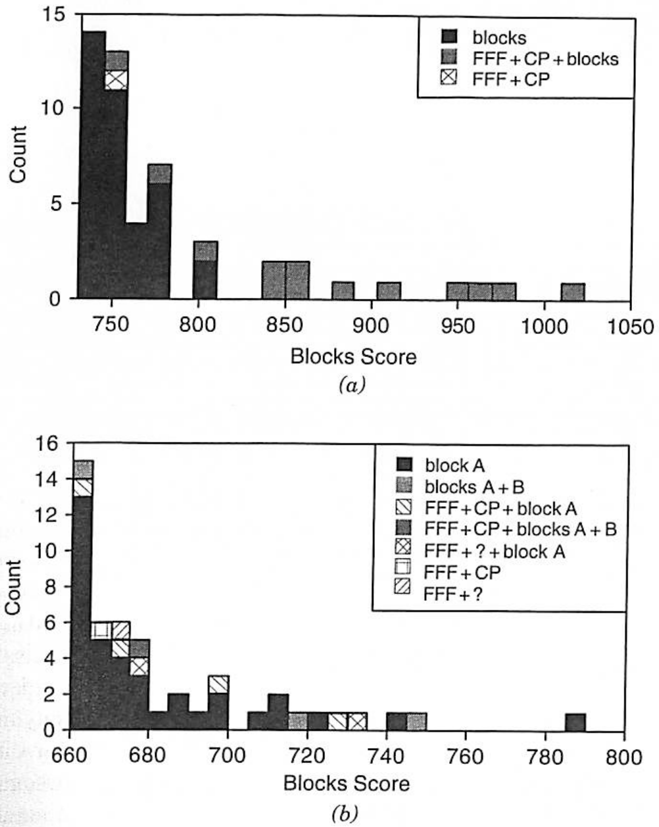


Figure 3. For the *B. subtilis* genome, the distribution of Blocks scores [236, 237] for the thioredoxin block and glutaredoxin blocks are presented. FFF indicates that the threaded structure satisfies the disulfide oxidoreductase active site descriptor, CP indicates that the sequence identified by threading and FFF satisfies the conservation profile, and ? indicates that there is just one sequence so that a CP analysis cannot be done.

would yield a significant number of false positives. Thus, what we require is a method that places such structures where their score is sufficiently significant that on subsequent filtration by a functional descriptor, they can be reliably identified. This is the origin of use of multiple scoring functions in PROSPECTOR1, which, in combination, selects 59 of 68 Fischer pairs in the top scoring position.

Surprisingly, despite the fact that threading algorithms have problems generating good sequence–structure alignments, we have found that active sites are often accurately aligned, even for very distant matches. This observation would agree with the above-mentioned experimental results that active sites are well-conserved in protein structures. Of course, because no genome has the function of all its proteins experimentally annotated, it is impossible to know how many proteins with the specified biochemical function are missed, nor is there yet experimental characterization of most of these predictions.

IX. USE OF LOW-RESOLUTION STRUCTURES FOR LIGAND IDENTIFICATION

One of the important elements of protein function is the ability of a protein to interact with and bind various ligands. This ability is closely related to the three-dimensional structure of the protein. Because the quality of theoretical structure prediction methods has recently improved considerably, we are developing a docking procedure that will utilize these relatively low-quality models of proteins for the prediction of plausible conformations of receptor-small ligand complexes as well as for the prediction of interactions between particular subunits of a protein in the quaternary structures.

Our approach to the problem of low-resolution docking focuses on the steric and quasi-chemical complementarity between the ligand and the receptor molecules. Because the predicted structures that result from theoretical predictions usually resemble very low-resolution experimental structures, in our method we use only approximate models of both the ligand and its receptor. Vakser et al. [240] have demonstrated that by averaging the structural details of interacting molecules it is possible to drive the docking procedure toward the real binding site, thus avoiding, in many cases, the local minima problem. It also turns out in our case that this averaging procedure allows for the compensation of the numerous structural inaccuracies that result from the theoretical predictions of the receptor structure.

In the first stage of our docking procedure, structures of both molecules, the receptor and the ligand, are projected onto a uniform cubic lattice, thus giving two clusters of adjacent cubes. These two clusters approximate the shapes of both molecules with the accuracy of the grid size. Some of the receptor cubes ("surface" cubes) can be penetrated by the ligand, leading to favorable interactions when overlapped with the ligand, whereas others (interior cubes) contribute to the repulsive contacts. As elegantly demonstrated by Vakser et al. [240], when such a procedure is correctly implemented, this simple steric matching protocol is often quite successful in rebuilding correctly docked complexes.

While the steric method described above is very efficient, in many cases, geometric criteria alone are insufficient to correctly dock the two molecules. This is especially true when the structure of the receptor is of poor quality or a ligand molecule is relatively small so that shape complementarity is insufficient to specify the correct conformation. To overcome this problem, we decided to build a statistical potential that could be used for additional evaluation of the quality of the match. In order to build the potential, we defined 20 general atom types and built the contact statistics on the basis of the structures of known complexes available in the PDB [171]. After projection of the two molecules onto the grid, every cube is additionally labeled with the properties defined by the atom types that were projected onto it. Once the approximate representation of the system is ready, the best match of these two cube-clusters is determined by exhaustive scanning over the six-dimensional conformational space of the three relative translations and the three rotations. Calculating the value of the correlation function between these two sets of cubes and the value of the potential function, the quality of the particular ligand-receptor orientation is scored.

We applied this algorithm to predict (actually postdict) the structures of several complexes available in the protein data bank. These complexes include members of the Fischer database that had co-crystallized ligands that were generated by the procedure that was described in Section V. In most cases, not only is the location of the binding site on the receptor surface correctly identified, but the proper orientation of the bound ligand was reasonably well recovered as well, within the level of accuracy of the modeled receptor itself. In many cases, even structures of receptors as far as 5–6 Å away from native turned out to be accurate enough for the docking procedure to succeed.

Table VIII below shows five examples of the homology-modeled structures that were used in our docking calculations. The quality of the modeled receptor

TABLE VIII
Results of Docking Ligands to Low-resolution Predicted Structures[a]

Structure Name	RMSD of the Receptor from Native	Relative Shift of the Ligand from Native
2sarA	5.99	3.1
2cmd_	5.57	1.3
1bbhA	3.16	1.6
1mdc_	4.92	2.6
1c2rA	4.94	3.3

[a] All dimensions are in angstroms.

(in RMSD) and shift of the docked ligand relative to its position in the superimposed native complex are also shown.

Two examples of docked ligands to the generalized homology modeled receptors are shown in Fig. 4. The red is the native orientation of the ligand, and the yellow is the best scoring match. As is immediately evident, the algorithm does a reasonably good job in docking the ligand to the correct binding site in the correct orientation. While our method is still under active development, it has already revealed its usefulness in the successful docking calculations of even small ligands to the theoretically modeled receptors. When complete, this methodology could hopefully be used for the large-scale screening of the potential ligands for the receptors predicted from genomic sequences.

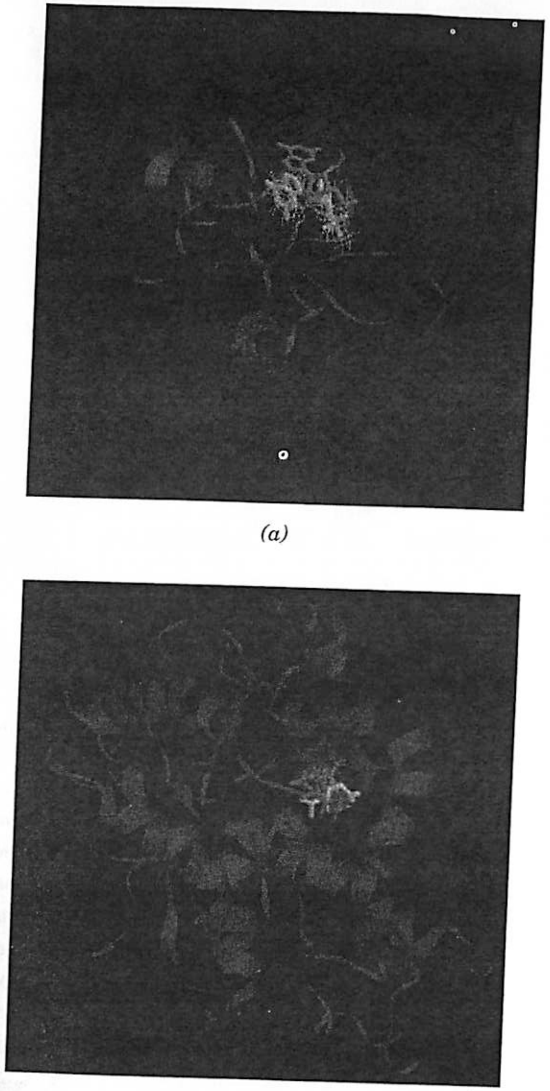
X. OUTLOOK FOR THE FUTURE

A. Possible Improvements of the Structure Prediction Methodology

The methodology for protein structure prediction outlined in this contribution, while partially successful, needs further improvement. First of all, some elements of the force field of the lattice model are not yet satisfactory. The threading algorithm PROSPECTOR, which forms the core of this approach, needs improvement. For example, it currently uses a very simple sequence profile, and more powerful techniques for generating more sensitive sequence profiles [241] need to be exploited. PROSPECTOR also generates high-scoring local sequence fragments that are often, but not always, quite accurate. This information needs to be incorporated into subsequent threading iterations as well as into partial seed structures in *ab initio* folding, akin to ROSETTA [242,243]. Better means of assessing the quality of the alignments also need to be developed.

The most promising way to improve generalized homology modeling is to couple the strength of template restraints to the quality of the template. Now, for all tested cases, the template-related restraints are of the same strength. Much better results may be possible if, for the templates that are close to the probe's structure, the restraints were very strong. For templates that are far from the probe's structure, the restraints should be very weak. The template should be used only for a loose definition of the fold topology. This requires an up-front estimation of the template quality in a semiquantitative fashion. Better scoring of the threading results and comparison with related cases (size of protein, percentage of alignment, comparison of the template alignments to other related proteins, etc.) might provide necessary data for the case-dependent scaling of the template-related restraints in the generalized homology modeling procedures.

Turning to issues associated with *ab initio folding* and, to a lesser extent, generalized comparative modeling, some elements of the force field of the lattice model are not yet satisfactory. The scaling of various contributions to the



(*a*)

(*b*)

Figure 4. (See also color insert.) For the predicted protein structure of 2sarA (2cmd_) generated by GeneComp using a template provided by the Fischer Database [34], the red-colored ligand represents the superposition of the ligand bound to the native receptor. The highest-scored match is colored in yellow.

interaction scheme is now to a large extent arbitrary and adjusted essentially by a trial-and-error method. A more precise scaling will be attempted by an automated procedure targeted to generating strong (as strong as possible) correlations between RMSD from correct folds and energy. A large set of decoys (lattice structures at various distances from native) will be used for this purpose. The weakest elements of the force field will be reexamined. Probably the largest improvement of the model could be achieved via introduction of approximate electrostatics into the interaction scheme. This should include more implicit treatment of the solvent and other than intra-main-chain hydrogen bonds.

For *ab initio* folding, a better means of the fold selection is needed. As mentioned above, for the majority of small proteins, the SICHO simulations produce a fraction of very good low-to-moderate resolution structures. Unfortunately, the model force field is capable of selecting these good folds in only a fraction of cases. Perhaps the folding simulations and the fold selection procedures should be separated in a more radical way. It appears to make sense that different force fields may be more efficient for folding simulations than those used for the fold selection. Indeed, folding requires an interaction scheme that discriminates not only against the wrong folds but also against a huge part of model-chain conformational space that does not correspond to any protein structures. The fold selection stage needs potentials that essentially discriminate between various protein-like conformations. Fortunately, fold selection involves a few hundred structures. Thus, more detailed, including all-atom, interaction schemes could be employed.

B. In Combination with Experiment

A variety of fragmentary experimental data could be used to increase the accuracy and to extend the range of applicability of the described methodology for protein structure prediction. The *ab initio* folding procedure employs predicted secondary structure (in a three-letter code) and predicted contacts between side groups. None of these predictions are exact; this has a consequence for the overall performance of the method. Knowledge of the exact protein secondary structure or some elements of secondary structure significantly increases the precision and accuracy of the three-dimensional structure predictions. Also, the exact knowledge of a few side-chain contacts increases the applicability of the method. As demonstrated recently [139] for an older version of the SICHO model, knowledge of secondary structure and as few as $N/7$ to $N/5$ side-chain contacts (where N is the number of residues in the protein) enable reproducible structure assembly for proteins up to 240 residues. The larger the number of known contacts, the better the accuracy of the predicted structures. Such fragmentary structural data could be extracted from NMR experiments. When more extensive data are difficult (or impossible) to obtain, the lattice folding provides a low-to-moderate resolution molecular model of the protein of interest. In those cases where a lot of NMR-based restraints are collected, the possibility of obtaining of an approximate model from just a few identified long-range contacts may aid with assignment processes for the other signals. Such a procedure can be iterated. Alternately, such constraints could be implemented in PROSPECTOR as a potential to help improve the quality of fold selection as well as the quality of alignments. Structural restraints for the *ab initio* folding can originate not only from NMR data but also from electron microscopy. Fluorescence data or crosslinking experiments could also provide some information about the side-chain contacts. Sometimes, mutation experiments can identify residues that are involved with ligand binding. Information about the spatial arrangement of these residues could be easily incorporated into the folding algorithm. Another type of possible connection with experiment is probably worth mentioning. Sometimes, as a result of *ab initio* folding simulations, not one but a few plausible folds are generated. When compared with experiments required for structure determination from scratch, a much simpler experiment could be designed and executed for the selection between a few possible structures.

C. Improvement of Structure-Based Biochemical Function Prediction

A key component of the ability to predict the biochemical function of a protein using a structure-based approach is the availability of an extensive active site library. Once this is available, then the assignment of biochemical function can be done with a far smaller false-positive rate than alternative sequence-based approaches [15,244]. While active site FFFs can be built by hand, such a process is very time consuming, and automated approaches to active site identification must be developed. One such approach used PDB descriptors to assign active site residues [14], but more recent work using conservation profile analysis of these site descriptors indicates a significant false-positive rate [245]. However, if the identified active site residues are conserved, then one can tentatively build a functional descriptor on this basis. Alternatively, one could use BLOCKS [236] to identify conserved positions and attempt to build a three-dimensional descriptor on a unique subset of highly conserved residues [246]. We are currently undertaking such an approach.

To date, no large-scale refinement of the alignments generated by threading has been undertaken. If the alignment is in error and active site residues are not correctly aligned, then a false negative will result. Thus, we plan to apply GeneComp to demonstrate the stability of correct alignments (i.e., to show that true positives do not become false negatives). Next we plan to test the method on the weakly significant alignments (Z score > 1) first for *M. genitalium* and then for *E. coli*. If our results on the Fischer database are a guide, not only will this provide a set of better models for a significant fraction of both genomes, but perhaps, using a more complete active site library, additional ORFans can be assigned.

D. Improvement of Low-to-Moderate Resolution Docking of Ligands

Thus far we have demonstrated that in roughly 50% of the cases, the binding conformation of a known ligand can be identified using a low-resolution (backbone RMSD from native up to about 6 Å) predicted structure. While these results are encouraging, much more must be done. The energetic description describing the interaction of ligand and receptor must be improved so that the accuracy of the method is enhanced, and systematic clustering of the results using our clustering algorithm [197] must be done. Moreover, it remains to be demonstrated that unknown ligands can be identified using such an approach. Even if it turns out that in a library of several hundred thousand to millions of compounds, one could only place true ligands in the 500th position or so (a realistic goal for a low-resolution model), this would be quite valuable. Future work is proceeding along these lines.

The low-resolution description could also be used to dock macromolecular complexes. We have had very encouraging preliminary results on correctly docking the dimer in the tobacco mosaic virus, but clearly much more thorough benchmarking is required. One might imagine predicting the tertiary structure of two molecules and then docking them, but such studies are in the very preliminary stage.

E. Summary

In this review, we have described a number of approaches to the prediction of protein structure and biochemical function. A key theme of this review is that low-to-moderate resolution structures by state-of-the-art techniques are quite valuable. If the structure has a backbone RMSD from native in the range of 4–6 Å, it can be used to identify the biochemical function of a protein, and known ligands can be docked to identify the binding site as well as a low-resolution prediction of the location of the ligand in the receptor. The question then is, What are contemporary techniques for low-resolution protein structure prediction? After having reviewed the state of the field, which includes a number of promising *ab initio* studies [128,133,141,142,146] and threading algorithms [39, 53–56], we then introduced a unified approach to protein structure prediction. This methodology involves the use of a newly developed, iterative threading algorithm, PROSPECTOR [57], where one threads first (see Fig. 1). If there is no significant match to a template structure, the consensus contacts and secondary structure in the top 20 scoring structures are used as restraints in an *ab initio* folding algorithm. On average, this contact prediction predicts about one-third of the contacts correctly and predicts above 70% correctly within two residues. Application of this methodology to a representative test set of 28 structures results in the native state (of low-resolution structures up to 6.5 Å) being in one of the well-defined clusters in 15 cases. If fold selection is done not in the reduced model but in an atomic model, then 17 cases are foldable. Conversely, if PROSPECTOR identifies a global template, then we perform generalized comparative modeling, GeneComp, to refine the structures. This procedure uses the template alignment, as well as predicted contacts and secondary structure (not necessarily from the template structure), as restraints. In practice, when applied to representative probe proteins in the Fischer database [34,179], GeneComp tends to perform better on average than Modeller [23,27]. Moreover, it does no harm, that is, the quality of the model is either left the same or improves. Thus, it can be used with impunity. As in *ab initio* folding, the resulting structures are clustered and representative folds selected.

PROSPECTOR itself has been used to predict the tertiary structures of the proteins in two genomes, *M. genitalium* and *E. coli*, and successfully matches about 40% of the sequences to a known fold. Application of the three-dimensional active site descriptors designed for low-resolution structures, FFFs [8,10], allows one to select all known true positives, even when the *Z* score is close to 1. Furthermore, threading followed by application of the FFF has a far smaller false-positive rate than alternative sequence-based approaches such as BLOCKS [236,246]. Such approaches need to be generalized from treating enzymes to more generalized binding and macromolecular recognition.

This review describes one such way to use low-resolution structures to identify the binding site and conformation when one has a known ligand. The methodology was applied to those probe structures in the Fischer database that co-crystallized with ligands. As shown in Table VIII, it is possible to identify the binding conformation with moderate accuracy, even when the backbone RMSD from native is 6 Å. This opens up the possibility of genome scale screening of low resolution predicted structures for ligand binding.

While considerable progress has been made, there are significant challenges remaining. The generalized comparative modeling approach, GeneComp, needs to be extended so that it can treat highly homologous as well as analogous structures. Furthermore, given that *ab initio* folding algorithms quite often generate native-like structures, as also seen in generalized comparative modeling, development of better protein representations and energy functions that can select native folds from misfolded states is more crucial than ever. Clustering helps to reduce the problem by selecting representative folds, but routine unequivocal selection of native-like structures is not yet possible. It seems that the most promising approach is to convert the reduced models to full-atom models and then use either physics or knowledge-based energy functions to select the native structure. Use of active site descriptors can also help in this regard, because they act like a filter. Because of their utility in biochemical function assignment, better techniques for the construction of functionally relevant active sites is a must. Finally, while considerable progress has been made in the docking of known small-molecule ligands to low-resolution structures, methods must be developed that can identify such ligands, at the least by enriching the yield of true positives. Work in this direction is underway.

In conclusion, while techniques for the prediction of low-resolution structures have improved, they still have a way to go before structure prediction becomes routine. Nevertheless, this is a very laudable goal because low-resolution structures are of considerable utility both in the identification of biochemical function and in ligand docking. Such efforts will have to be applied on a genomic scale if structure-based approaches to function prediction are to play a role in the post genomic era. A number of such efforts are underway, and doubtless there will be more in the future.

Acknowledgment

This research was supported in part by NIH grants Nos. GM37408, GM48835, and RR-12255 and NSF grant No. 9986019. The contributions of Drs. Betancourt, Fetrow, Kihara, Ilkowski, Rotkiewicz, Wojciechowski, and N. Siew to some of the research described in this proposal are greatly appreciated. Andrzej Kolinski also acknowledges support by the University of Warsaw, grant No. 120-501/68-BW-1483/18/2000.

References

1. S. F. Altschul, W. Gish, W. Miller, et al., *J. Mol. Biol.* **215**, 403 (1990).
2. W. R. Pearson, *Methods Enzymol.* **266**, 227 (1996).
3. A. Bairoch, P. Bucher, and K. Hofmann, *Nucleic Acids Res.* **24**, 189 (1995).
4. S. Henikoff and J. G. Henikoff, *Genomics* **19**, 97 (1994).
5. T. K. Attwood, M. E. Beck, A. J. Bleasby, et al., *Nucleic Acids Res.* **22**, 3590 (1994).
6. T. K. Attwood, M. E. Beck, A. J. Bleasby, et al., *Nucleic Acids Res.* **25**, 212 (1997).
7. C. G. Nevill-Manning, T. D. Wu, and D. L. Brutlag, *Proc. Natl. Acad. Sci. USA* **95**, 5865 (1998).
8. J. S. Fetrow and J. Skolnick, *J. Mol. Biol.* **281**, 949 (1998).
9. L. Yu, J. V. White, and T. F. Smith, *Protein Sci.* **7**, 2499 (1998).
10. J. S. Fetrow, A. Godzik, and J. Skolnick, *J. Mol. Biol.* **282**, 703 (1998).
11. L. Zhang, A. Godzik, and J. Skolnick, et al., *Fold. Des.* **3**, 535 (1998).
12. J. S. Fetrow, N. Siew, and J. Skolnick, *FASEB J.* **13**, 1866 (1999).
13. N. Siew, J. Skolnick, and J. Fetrow, in preparation (2000).
14. B. Zhang, L. Rychlewski, and K. Pawlowski, et al., *Protein Sci.* **8**, 1104 (1999).
15. J. Skolnick and J. Fetrow, *TIBTECH* **18**, 34 (2000).
16. Z. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **84**, 6611 (1987).
17. J. Moult, T. Hubbard, K. Fidelis, et al., *Proteins Suppl.* **2** (1999).
18. P. A. Bates and M. J. Sternberg, *Proteins Suppl.* **47** (1999).
19. D. F. Burke, C. M. Deane, H. A. Nagarajaram, et al., *Proteins Suppl.* **55** (1999).
20. A. S. Yang and B. Honig, *Proteins Suppl.* **66** (1999).
21. R. L. Dunbrack, Jr., *Proteins Suppl.* **81** (1999).

22. D. Fischer, *Proteins Suppl.* **61** (1999).
23. R. Sanchez and A. Sali, *Proteins Suppl.* **50** (1997).
24. A. Sali, L. Potterton, F. Yuan, et al., *Proteins* **23**, 318 (1995).
25. T. Alwyn Jones and G. J. Kleywegt, *Proteins Suppl.* **30** (1999).
26. A. Zemla, C. Venclovas, and J. Moult, et al., *Proteins Suppl.* **22** (1999).
27. R. Sanchez and A. Sali, *Proc. Natl. Acad. Sci. USA* **95**, 13597 (1998).
28. A. Goffeau, B. G. Barrell, H. Bussey, et al., *Science* **274**, 546 (1996).
29. A. Elofsson and E. L. Sonnhammer, *Bioinformatics* **15**, 480 (1999).
30. R. Lathrop and T. F. Smith, *J. Mol. Biol.* **255**, 641 (1996).
31. R. T. Miller, D. T. Jones, and J. M. Thornton, *FASEB* **10**, 171 (1996).
32. A. Kolinski, P. Rotkiewicz, B. Ilkowski, et al., *Proteins* **37**, 592 (1999).
33. M. Wilmanns and D. Eisenberg, *Proc. Natl. Acad. Sci. USA* **90**, 1379 (1993).
34. D. Fischer, A. Elofsson, D. Rice, et al., *Pac. Symp. Biocomput.* **300** (1996).
35. A. R. Panchenko, A. Marchler-Bauer, and S. H. Bryant, *J. Mol. Biol.* **296**, 1319 (2000).
36. T.-M. Yi and E. S. Lander, *Protein Sci.* **3**, 1315 (1994).
37. Y. Matsuo and K. Nishikawa, *Protein Sci.* **3**, 2055 (1994).
38. D. T. Jones, *J. Mol. Biol.* **292**, 195 (1999).
39. K. K. Koretke, R. B. Russell, R. R. Copley, et al., *Proteins Suppl.* **141** (1999).
40. V. N. Maiorov and G. M. Crippen, *J. Mol. Biol.* **277**, 876 (1992).
41. A. Tropsha, R. K. Singh, I. I. Vaisman, et al., in *Pacific Symposium on Biocomputing '96*, L. Hunter and T. E. Klein, eds. World Scientific, Singapore, 1996, p. 614.
42. D. T. Jones, W. R. Taylor, and J. M. Thornton, *Nature* **358**, 86 (1992).
43. K. K. Koretke, Z. Luthey-Schulten, and P. G. Wolynes, *Protein Sci.* **5**, 1043 (1996).
44. S. H. Bryant and C. E. Lawrence, *Proteins* **16**, 92 (1993).
45. A. Godzik, J. Skolnick, and A. Kolinski, *J. Mol. Biol.* **227**, 227 (1992).
46. J. Selbig, *Protein Eng.* **8**, 339 (1995).
47. M. J. Sippl and S. Weitckus, *Proteins* **13**, 258 (1992).
48. M. Wilmanns and D. Eisenberg, *Protein Eng.* **8**, 626 (1995).
49. S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.* **48**, 443 (1970).
50. J. U. Bowie, R. Luthy, and D. Eisenberg, *Science* **253**, 164 (1991).
51. R. Thiele, R. Zimmer, and T. Lengauer, *ISMB* **3**, 384 (1995).
52. A. G. Murzin, *Proteins* **37**, 88 (1999).
53. D. T. Jones, M. Tress, K. Bryson, et al., *Proteins Suppl.* **104** (1999).
54. M. Ota, T. Kawabata, A. R. Kinjo, et al., *Proteins Suppl.* **126** (1999).
55. F. S. Domingues, W. A. Koppensteiner, M. Jaritz, et al., *Proteins Suppl.* **112** (1999).
56. A. Panchenko, A. Marchler-Bauer, and S. H. Bryant, *Proteins Suppl.* **133** (1999).
57. J. Skolnick and D. Kihara, *Proteins*, in press (2000).
58. M. Levitt and A. Warshel, *Nature* **253**, 694 (1975).
59. M. Levitt, *J. Mol. Biol.* **104**, 59 (1976).
60. I. D. Kuntz, *J. Am. Chem. Soc.* **97**, 4362 (1975).
61. I. D. Kuntz, G. M. Crippen, P. A. Kollman, et al., *J. Mol. Biol.* **106**, 983 (1976).
62. A. T. Hagler and B. Honig, *Proc. Natl. Acad. Sci. USA* **75**, 554 (1978).

63. C. Wilson and S. Doniach, *Proteins* **6**, 193 (1989).
64. S. Sun, *Protein Sci.* **2**, 762 (1993).
65. A. Wallqvist and M. Ullner, *Proteins* **18**, 267 (1994).
66. D. Hoffmann and E. W. Knapp, *Eur. Biophys. J.* **24**, 387 (1996).
67. D. Hoffmann and E. W. Knapp, *Phys. Rev. E* **53**, 4221 (1996).
68. J. T. Pedersen and J. Moult, *Proteins Suppl.* **179**, 1 (1997).
69. N. Gō and H. Taketomi, *Proc. Natl. Acad. Sci. USA* **75**, 559 (1978).
70. W. R. Krigbaum and A. Komoriya, *Biochim. Biophys. Acta* **576**, 204 (1979).
71. W. R. Krigbaum and S. F. Lin, *Macromolecules* **15**, 1135 (1982).
72. A. Kolinski and J. Skolnick, *Proc. Natl. Acad. Sci. USA* **83**, 7267 (1986).
73. A. Kolinski, J. Skolnick, and R. Yaris, *J. Chem. Phys.* **85**, 3585 (1986).
74. A. Kolinski, J. Skolnick, and R. Yaris, *Biopolymers* **26**, 937 (1987).
75. J. Skolnick, A. Kolinski, and R. Yaris, *Proc. Natl. Acad. Sci. USA* **85**, 5057 (1988).
76. J. Skolnick and A. Kolinski, *Annu. Rev. Phys. Chem.* **40**, 207 (1989).
77. J. Skolnick, A. Kolinski, and R. Yaris, *Biopolymers* **28**, 1059 (1989).
78. J. Skolnick, A. Kolinski, and R. Yaris, *Proc. Natl. Acad. Sci. USA* **86**, 1229 (1989).
79. J. Skolnick and A. Kolinski, *J. Mol. Biol.* **212**, 787 (1990).
80. J. Skolnick and A. Kolinski, *Science* **250**, 1121 (1990).
81. A. Kolinski, M. Milik, and *J. Skolnick, J. Chem. Phys.* **94**, 3978 (1991).
82. A. Kolinski and J. Skolnick, *J. Phys. Chem.* **97**, 9412 (1992).
83. A. Kolinski and J. Skolnick, *Proteins* **18**, 338 (1994).
84. A. Kolinski, M. Milik, J. Rycombel, et al., *J. Chem. Phys.* **103**, 4312 (1995).
85. A. Sikorski and J. Skolnick, *Biopolymers* **28**, 1097 (1989).
86. A. Sikorski and J. Skolnick, *Proc. Natl. Acad. Sci. USA* **86**, 2668 (1989).
87. A. Sikorski and J. Skolnick, *J. Mol. Biol.* **215**, 183 (1990).
88. A. Sikorski and J. Skolnick, *J. Mol. Biol.* **212**, 819 (1990).
89. H. S. Chan and K. A. Dill, *J. Chem. Phys.* **92**, 492 (1989).
90. H. S. Chan and K. A. Dill, *Macromolecules* **22**, 4559 (1989).
91. H. S. Chan and K. A. Dill, *Proc. Natl. Acad. Sci. USA* **87**, 6388 (1990).
92. H. S. Chan and K. A. Dill, *Annu. Rev. Biophys. Biophys. Chem.* **20**, 447 (1991).
93. K. A. Dill, *Biochemistry* **24**, 1501 (1985).
94. K. A. Dill, D. O. V. Alonso, and K. Hutchinson, *Biochemistry* **28**, 5439 (1989).
95. K. A. Dill, *Curr. Biol.* **3**, 99 (1993).
96. K. A. Dill, S. Bromberg, K. Yue, et al., *Protein Sci.* **4**, 561 (1995).
97. A. Sali, E. Shakhnovich, and M. Karplus, *J. Mol. Biol.* **235**, 1614 (1994).
98. A. Sali, E. Shakhnovich, and M. Karplus, *Nature* **369**, 248 (1994).
99. E. I. Shakhnovich and A. V. Finkelstein, *Biopolymers* **28**, 1667 (1989).
100. E. I. Shakhnovich and A. M. Gutin, *Biophys. Chem.* **34**, 187 (1989).
101. E. I. Shakhnovich, G. Farztdinov, and A. M. Gutin, *Phys. Rev. Lett.* **67**, 1665 (1991).
102. E. I. Shakhnovich and A. M. Gutin, *Proc. Natl. Acad. Sci. USA* **90**, 7195 (1993).
103. E. I. Shakhnovich and A. M. Gutin, *Protein Eng.* **6**, 793 (1993).
104. E. I. Shakhnovich, *Phys. Rev. Lett.* **72**, 3907 (1994).

105. E. I. Shakhnovich, *Fold. Des.* **1**, R50 (1996).
106. A. R. Dinner, A. Sali, M. Karplus, et al., *J. Chem. Phys.* **101**, 1444 (1994).
107. A. R. Dinner, A. Sali, and M. Karplus, *Proc. Natl. Acad. Sci. USA* **93**, 8356 (1996).
108. M.-H. Hao and H. A. Scheraga, *J. Phys. Chem.* **98**, 4940 (1994).
109. M.-H. Hao and H. A. Scheraga, *J. Phys. Chem.* **98**, 9882 (1994).
110. M.-H. Hao and H. A. Scheraga, *J. Chem. Phys.* **102**, 1334 (1995).
111. A. Kolinski and P. Madziar, *Biopolymers* **42,** 537 (1997).
112. V. G. Dashevskii, *Mol. Biol.* (translation from) **14**, 105 (1980).
113. D. G. Covell, *Proteins* **14**, 409 (1992).
114. D. G. Covell and R. L. Jernigan, *Biochemistry* **29**, 3287 (1990).
115. D. A. Hinds and M. Levitt, *Proc. Natl. Acad. Sci. USA* **89**, 2536 (1992).
116. A. Kolinski and J. Skolnick, *Lattice Models of Protein Folding, Dynamics and Thermodynamics*, R. G. Landes, Austin, TX, 1996.
117. J. Skolnick and A. Kolinski, *J. Mol. Biol.* **221**, 499 (1991).
118. J. Skolnick, A. Kolinski, C. Brooks III, et al., *Curr. Biol.* **3**, 414 (1993).
119. A. Kolinski and J. Skolnick, *Acta Biochim. Polon.* **44**, 389 (1998).
120. A. Kolinski, P. Rotkiewicz, and J. Skolnick, in *Monte Carlo Approaches to Biopolymers and Protein Folding*, P. Grassberger, G. T. Barkema, and W. Nadler, eds., World Scientific, Singapore, 1998, p. 110.
121. A. Kolinski, A. Godzik, and J. Skolnick, *J. Chem. Phys.* **98**, 7420 (1993).
122. A. Kolinski and J. Skolnick, *Proteins* **18**, 353 (1994).
123. A. Kolinski and J. Skolnick, *Proteins* **18**, 338 (1994).
124. M. Vieth, A. Kolinski, C. L. Brooks III, et al., *J. Mol. Biol.* **1994**, 361 (1994).
125. M. Vieth, A. Kolinski, I. Brooks, C. L., et al., *J. Mol. Biol.* **251**, 448 (1995).
126. M. Vieth, A. Kolinski, and J. Skolnick, *Biochemistry* **35**, 955 (1996).
127. J. Moult, T. Hubbard, K. Fidelis, et al., *Proteins Suppl.* **3**, 2 (1999).
128. K. T. Simons, R. Bonneau, I. Ruczinski, et al., *Proteins Suppl.* **3**, 171 (1999).
129. B. Rost and C. Sander, *J. Mol. Biol.* **232**, 584 (1993).
130. B. Rost and C. Sander, *Proteins* **19**, 55 (1994).
131. B. Rost and C. Sander, *Proteins* **23**, 295 (1996).
132. D. T. Jones, *Proteins Suppl.* **185** (1997).
133. A. R. Ortiz, A. Kolinski, P. Rotkiewicz, et al., *Proteins Suppl.* **3**, 177 (1999).
134. D. Kihara, H. Lui, A. Kolinski, and J. Skolnick, *Proc. Natl. Acad. Sci. USA* **14**, 14 (2001).
135. A. R. Ortiz, W.-P. Hu, A. Kolinski, et al., in *Pacific Symposium on Biocomputing '97*, R. B. Altman, A. K. Dunker, L. Hunter, and T. E. Klein, eds., World Scientific, Singapore, 1997, p. 316.
136. A. R. Ortiz, A. Kolinski, and J. Skolnick, *Proc. Natl. Acad. Sci. USA* **95**, 1020 (1998).
137. A. R. Ortiz, A. Kolinski, and J. Skolnick, *J. Mol. Biol.* **277**, 419 (1998).
138. A. R. Ortiz, A. Kolinski, and J. Skolnick, *Proteins* **30**, 287 (1998).
139. A. Kolinski and S. A., *Proteins* **32**, 475 (1998).
140. A. Kolinski, P. Rotkiewicz, B. Ilkowski, et al., *Progress of Theoretical Physics (Kyoto) Suppl.* **138**, 292 (2000).
141. D. J. Osguthorpe, *Proteins Suppl.* **3**, 186 (1999).

142. R. Samudrala, H. Xia, E. Huang, et al., *Proteins Suppl.* **3**, 194 (1999).
143. D. Hinds and M. Levitt, *J. Mol. Biol.* **243**, 668 (1994).
144. B. Park, E. Huang, and M. Levitt, *J. Mol. Biol.* **266**, 831 (1997).
145. E. Huang, R. Samudrala, and J. Ponder, *Protein Sci.* **7**, 1998 (1998).
146. J. Lee, A. Liwo, D. R. Ripoll, et al., *Proteins Suppl.* **3**, 204 (1999).
147. J. Lee, A. Liwo, and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **96**, 2025 (1999).
148. A. Liwo, R. Kazimierkiewicz, C. Czaplewski, et al., *J. Comput. Chem.* **19**, 259 (1988).
149. D. R. Ripoll, A. Liwo, and H. A. Scheraga, *Biopolymers* **46**, 117 (1988).
150. A. Baumgaertner, in *The Monte Carlo Method in Condensed Matter Physics*, K. Binder, ed., Springer, Heidelberg, 1995.
151. A. Rey and J. Skolnick, *Chem. Phys.* **158**, 199 (1991).
152. J. Skolnick and A. Kolinski, in *Computer Simulations of Biomolecular Systems. Theoretical and Experimental Studies*, W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, eds., ESCOM Science Publishers, 1996.
153. A. Kolinski, B. Ilkowski, and J. Skolnick, *Biophys. J.* **77**, 2942 (1999).
154. U. H. E. Hansmann and Y. Okamoto, *J. Comput. Chem.* **14**, 1333 (1993).
155. U. H. E. Hansmann and Y. Okamoto, *J. Chem. Phys.* **110**, 1267 (1999).
156. A. Kolinski, W. Galazka, and J. Skolnick, *Proteins* **26**, 271 (1996).
157. A. Kolinski, W. Galazka, and J. Skolnick, *J. Chem. Phys.* **108**, 2608 (1998).
158. C. B. Anfinsen, *Science* **181**, 223 (1973).
159. H. A. Scheraga, *Biophys. Chem.* **59**, 329 (1996).
160. L. Piela, J. Kostrowicki, and H. A. Scheraga, *J. Phys. Chem.* **93**, 3339 (1989).
161. H. A. Scheraga and M.-H. Hao, *Adv. Chem. Phys.* **105**, 243 (1999).
162. Z. Li and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **84**, 6611 (1987).
163. A. A. Rabow and H. A. Scheraga, *Protein Sci.* **5**, 1800 (1996).
164. T. Dandekar and P. Argos, *J. Mol. Biol.* **256**, 645 (1996).
165. Z. Sun, X. Xia, Q. Guo, et al., *J. Protein Chem.* **18**, 39 (1999).
166. U. H. E. Hansmann and Y. Okamoto, *J. Comput. Chem.* **18**, 920 (1997).
167. U. H. E. Hansmann and Y. Okamoto, *Curr. Opin. Struct. Biol.* **9**, 177 (1999).
168. D. Gront, A. Kolinski, and J. Skolnick, *J. Chem. Phys.* **113**, 5065 (2000).
169. R. H. Swedensen and J. S. Wang, *Phys. Rev. Lett.* **57**, 2607 (1986).
170. S. F. Altschul and E. V. Koonin, *Trends Biochem Sci.* **23**, 444 (1998).
171. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, et al., *J. Mol. Biol.* **112**, 535 (1977).
172. J. G. Henikoff and S. Henikoff, *Methods Enzymol.* **266**, 88 (1996).
173. A. Ogiwara, I. Uchiyama, T. Takagi, et al., *Protein Sci.* **5**, 1991 (1996).
174. C. Ouzounis, C. Sander, M. Scharf, et al., *J. Mol. Biol.* **232**, 805 (1993).
175. W. R. Pearson, *Methods Mol. Biol.* **24**, 307 (1994).
176. W. R. Pearson, *J. Mol. Biol.* **276**, 71 (1998).
177. J. D. Thompson, D. G. Higgins, and T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
178. J. Skolnick, A. Kolinski, and A. Ortiz, *Proteins* **38**, 3 (2000).
179. UCLA, http://www.doembi.ucla.edu/people/fischer/BENCH/table1.html, Los Angeles, 19
180. M. S. Waterman and M. Eggert, *J. Mol. Biol.* **197**, 723 (1987).

181. L. Jaroszewski, L. Rychlewski, B. Zhang, et al., *Protein Sci.* **7**, 1431 (1998).
182. S. F. Altschul, T. L. Madden, A. A. Schaffer, et al., *Nucleic Acids Res.* **25**, 3389 (1997).
183. J. Skolnick, L. Jaroszewski, A. Kolinski, et al., *Protein Sci.* **6**, 676 (1997).
184. C. M. Fraser, J. D. Gocayne, O. White, et al., *Science* **270**, 397 (1995).
185. D. Fischer and D. Eisenberg, *Proc. Natl. Acad. Sci. USA* **94**, 11929 (1997).
186. S. A. Teichmann, C. Chothia, and M. Gerstein, *Curr. Opin. Struct. Biol.* **9**, 390 (1999).
187. M. Gerstein, *Proteins* **33**, 518 (1998).
188. D. T. Jones, *J. Mol. Biol.* **287**, 797 (1999).
189. F. R. Blattner, G. Plunkett, 3rd, C. A. Bloch, et al., *Science* **277**, 1453 (1997).
190. L. Rychlewski, B. Zhang, and A. Godzik, *Protein Sci.* **8**, 614 (1999).
191. A. R. Ortiz, A. Kolinski, and J. Skolnick, *J. Mol. Biol.* **277**, 419 (1998).
192. A. R. Ortiz and Skolnick, *Biophys. J.* **79**, 1787 (2000).
193. A. Ortiz, A. Kolinski, P. Rotkiewicz, et al., *Proteins Suppl.* **3**, 177 (1999).
194. A. Aszodi and W. R. Tylor, *Fold. Des.* **1**, 325 (1996).
195. A. Sali and T. L. Blundel, *J. Mol. Biol.* **234**, 779 (1993).
196. L. Jaroszewski, K. Pawlowski, and A. Godzik, *J. Mol. Modelling* **00**, 000 (1998).
197. M. Betancourt and J. Skolnick, *J. Comput. Chem.* **22**, 339 (2001).
198. A. Kolinski, L. Jaroszewski, P. Rotkiewicz, et al., *J. Phys. Chem.* **102**, 4628 (1998).
199. A. Kolinski, P. Rotkiewicz, B. Ilkowski, et al., *Proteins* **37**, 592 (1999).
200. M. Feig, P. Rotkiewicz, A. Kolinski, et al., *Proteins* **41**, 86 (2000).
201. E. S. Huang, P. Koehl, M. Levitt, et al., *Proteins* **33**, 204 (1998).
202. J. Skolnick, A. Kolinski, and A. R. Ortiz, *Proteins* **38**, 3 (2000).
203. A. Sali, L. Potterton, F. Yuan, H. van Vlijmen, and M. Karplus, Evaluation of comparative protein modeling by MODELLER, *Proteins* **23**, 318–326 (1995).
204. H. Lu and J. Skolnick, *Proteins* **44**, 223 (2001).
205. W. Kabsch and C. Sander, *Biopolymers* **22**, 2577 (1983).
206. C. Chothia and A. V. Finkelstein, *Annu. Rev. Biochem.* **59**, 1007 (1990).
207. B. R. Brooks, R. Bruccoleri, B. Olafson, et al., *J. Comput. Chem.* **4**, 187 (1983).
208. C. Simmerling, M. Lee, A. R. Ortiz, et al., *J. Am. Chem. Soc.* **122**, 8392 (2000).
209. D. Mohanty, A. Kolinski, and J. Skolnick, *Biophys. J.* **77**, 54 (1999).
210. D. Mohanty, B. N. Dominy, A. Kolinski, et al., *Proteins* **35**, 447 (1999).
211. A. G. Murzin, S. E. Brenner, T. Hubbard, et al., *J. Mol. Biol.* **247**, 536 (1995).
212. C. A. Orengo, A. D. Michie, S. Jones, et al., *Structure* **5**, 1093 (1997).
213. A. M. Lesk, C.-I. Branden, and C. Chothia, *Proteins* **5**, 139 (1989).
214. G. K. Farber and G. A. Petsko, *Trends Biochem. Sci.* **15**, 228 (1990).
215. H. Hegyi and M. Gerstein, *J. Mol. Biol.* **288**, 147 (1999).
216. D. Fischer, H. Wolfson, S. L. Lin, et al., *Protein Sci.* **3**, 769 (1994).
217. A. Kasuya and J. M. Thornton, *J. Mol. Biol.* **286**, 1673 (1999).
218. C. D. Coldren, H. W. Hellinga, and J. P. Caradonna, *Proc. Natl. Acad. Sci. USA* **94**, 6635 (1997).
219. A. L. Pinto, H. W. Hellinga, and J. P. Caradonna, *Proc. Natl. Acad. Sci. USA* **94**, 5562 (1997).
220. H. W. Hellinga and F. M. Richards, *J. Mol. Biol.* **222**, 763 (1991).
221. H. W. Hellinga, J. P. Caradonna, and F. M. Richards, *J. Mol. Biol.* **222**, 787 (1991).

222. M. Klemba and L. Regan, *Biochemistry* **34**, 10094 (1995).
223. M. Klemba, K. H. Gardner, S. Marino, et al., *Nature Struct. Biol.* **2**, 368 (1995).
224. E. Farinas and L. Regan, *Protein Sci.* **7**, 1939 (1998).
225. M. W. Crowder, J. D. Stewart, V. A. Roberts, et al., *J. Am. Chem. Soc.* **117**, 5627 (1995).
226. S. Halfon and C. S. Craik, *J. Am. Chem. Soc.* **118**, 1227 (1996).
227. A. C. Wallace, R. A. Laskowski, and J. M. Thornton, *Protein Sci.* **5**, 1001 (1996).
228. G. J. Kleywegt, *J. Mol. Biol.* **285**, 1887 (1999).
229. A. C. Wallace, N. Birkakoti, and J. M. Thornton, *Protein Sci.* **6**, 2308 (1997).
230. R. B. Russell, *J. Mol. Biol.* **279**, 1211 (1998).
231. K. F. Han, C. Bystroff, and D. Baker, *Protein Sci.* **6**, 1587 (1997).
232. P. J. Artymiuk, A. R. Poirrette, H. M. Grindley, et al., *J. Mol. Biol.* **236**, 327 (1994).
233. S. Karlin and Z. Y. Zhu, *Proc. Natl. Acad. Sci. USA* **93**, 8344 (1996).
234. J. Fetrow, personal communication (2000).
235. E. E. Abola, F. C. Bernstein, S. H. Bryant, et al., *Protein Data Bank in Crystallographic Databases—Information Content, Software Systems, Scientific Application*, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987.
236. S. Henikoff, J. G. Henikoff, and S. Pietrokovski, *Bioinformatics* **15**, 471 (1999).
237. J. G. Henikoff, S. Henikoff, and S. Pietrokovski, *Nucleic Acids Res.* **27**, 226 (1999).
238. F. R. Blattner, G. Plunkett, 3rd, C. A. Bloch, et al., *Science* **277**, 1453 (1997).
239. F. Kunst and N. Ogasawara and I. Moszer, et al., *Nature* **390**, 249 (1997).
240. I. A. Vakser, O. G. Matar, and C. F. Lam, *Proc. Natl. Acad. Sci. USA* **96**, 8477 (1999).
241. L. Rychlewski, L. Jaroszewski, W. Li, et al., *Protein Sci.* **9**, 232 (2000).
242. C. Bystroff and D. Baker, *J. Mol. Biol.* **281**, 565 (1998).
243. K. T. Simons, R. Bonneau, I. Ruczinski, et al., *Proteins Suppl.* **171** (1999).
244. J. Skolnick, J. S. Fetrow, and A. Kolinski, *Nature Biotech.* **18**, 283 (2000).
245. T. Chiu and J. Skolnick, unpublished results (2000).
246. J. G. Henikoff, S. Pietrokovski, C. M. McCallum, et al., *Electrophoresis* **21**, 1700 (2000).
247. A. A. Schaffer, Y. I. Wolf, C. P. Ponting, et al., *Bioinformatics* **15**, 1000 (1999).