J. Leszczynski (Editor)
Computational Molecular Biology
Theoretical Computational Chemistry, Vol. 8

Chapter 11

Application of reduced models to protein structure prediction

J. Skolnick[a], A. Kolinski[a,b] and A. R. Ortiz[a]

[a]Department of Molecular Biology, The Scripps Research Institute,
10550 N. Torrey Pines Road, La Jolla, CA 92037 USA

[b]Department of Chemistry, University of Warsaw,
Pasteura 1, 02-093 Warsaw, Poland

1. INTRODUCTION

Each day the various plant, bacterial, Archea and eukaryotic genome sequencing projects generate additional protein sequence information at an ever increasing rate [1-10]. These raw data, being devoid of corresponding information about protein structure or function, are in and of themselves of extremely limited use [11]. To address the crucial problem of utilizing these data in the post genomic era, a means of predicting protein structure and/or function from sequence is required [11, 12]. To date, the most prevalent and successful methods of protein structure and function prediction are purely sequence based [13-15]. Unfortunately, these methods, which also include local sequence motif identification [16-18], are limited by the extent of sequence similarity between sequences of known and unknown proteins; they increasingly fail as the sequence identity diverges into and beyond the twilight zone of sequence identity between two proteins, which is about 30% [19]. In practice, roughly half of a given genome falls into this category [20, 21]. Alternatively, one might attempt to predict the protein's structure first, and then deduce from it the protein's function [12, 22, 23]. The latter approach is much more difficult because it is not necessarily based on evolutionary relationships and is still in its infancy. Nevertheless, a key component of

structure based approaches to function prediction is the ability to predict protein structure from sequence. Thus, in this review, we describe the state of the art of contemporary approaches to protein tertiary structure prediction, and focus in particular on reduced models.

1.1. Energy Functions and Search Protocols

Any successful tertiary structure prediction algorithm must address two intertwined issues: First, one must have an energy or fitness function that distinguishes the native conformation from the sea of alternative structures, which is in principle exponentially large. These energy functions might be based on first principles (e.g., from fitting to IR data on small molecules or from quantum mechanics [24-26]). Alternatively, they might incorporate knowledge about the general and specific features of proteins [27-30] (e.g. hydrophobic residues prefer to be buried and protein structures have almost all their hydrogen bonds satisfied), or might be a combination of the two [31]. Second, one must have a conformational search protocol that can find the native conformation among the possible alternative structures. A large variety of search schemes have been employed including molecular and Brownian dynamics simulations [32, 33], classical Metropolis Monte Carlo [34-41], entropy sampling Monte Carlo [42-45], the diffusion equation method that deforms the energy landscape [46, 47], and genetic algorithms [48, 49]. Currently, these are very active areas of investigation.

1.2. Protein representation

A key issue that one faces when embarking on a program of protein structure prediction is deciding on the level of detail of protein representation. At one extreme, all atoms in the protein including hydrogens are included. The motivation behind treating the system in such great detail is the hope that this geometric fidelity will translate into prediction accuracy [50-56]. However, such calculations are computationally very expensive and even for proteins on the order of 100 residues or so, at present they are impractical. While very encouraging results in a single Molecular Dynamics simulation of the folding of a small protein, the 36-residue villin headpiece, have been recently reported [57], at present such calculations are the exception rather than the rule. Furthermore, to ensure reproducibility, multiple independent folding simulations need to be done. Thus, in the interests of computational practicality, simplified protein models have been developed both to explore the general issues of protein folding thermodynamics and kinetics as well as prediction of tertiary structure of real proteins [27, 28, 39, 42, 43, 45, 58

In practice, the protein might simply be modeled as a set of described by the Cα backbone atom positions [50, 65, 66, 75, 7 representation may be further idealized so that the geometric fidelity proteins is very limited, as is the case in cubic lattice protein models An advantage of reduced models is their computational tractability, w disadvantage is that they have will have limited geometric accuracy a not be able to address certain questions that depend on atomic deta However, even if such atomic detail is the ultimate objective, reduced could form the basis of a hierarchical approach to structure prediction low to moderate resolution model is generated first [63, 83, 84]. Then, detail is added, and the resulting model is refined to the appropriate resolution. Such combined approaches retain the advantages of both and reduced protein models and if successful would not suffer fro respective disadvantages. The first step towards this goal dema development of computational methods that can predict the approximat structure. Here, a major emphasis is on the use of reduced protein m achieve this.

Given a particular choice of reduced protein model, one has to whether or not to describe a protein in continuous space or in a representation [33]. The advantage of a lattice is purely comput Because the protein is confined to a set of grid points, one can preca many geometric and energetic properties in advance. Thus, a well-d lattice model is about a factor of 10 to 100 times faster than the corres continuous space model [33, 85]. If the folding of a protein on the requires one CPU day, adequate conformational sampling in the con space model will not be practical. However, a lattice model also intro number of effects due to spatial anisotropy. For example, in ve coordination number lattices along certain lattice directions, the best ge representation of an α-helix may be achiral [58]. This is not a pa problem if general aspects of protein folding are to be investigated, bu be very problematic if one wants to predict the tertiary structure o proteins. On lattices of intermediate coordination, the best representation of a protein may be different along different directions, energy may change as one rotates the protein structure with respec lattice [85]. If such energetic changes are small, they are not too wo

but they could preclude selection of the native fold when they are substantial. Furthermore, assembled elements of secondary structure might be effectively frozen in space as the collective motion required to move a helix without hitting any other element in a compact structure might not be possible. As one goes to high coordination number lattices, these effects can be minimized, and in the very latest high coordination lattice models, they are essentially eliminated [86].

1.3. Use of simplified models to obtain general insights into protein folding

Obviously, even for reduced models, the requisite level of detail changes if one wishes to obtain insights into the general aspects of protein folding rather than in predicting the tertiary structure of a specific protein. A simple model where the exact enumeration of all compact states in short chains [87] can be done is the square or cubic lattice HP model [77, 88-90]. The polypeptide is represented as a string of hydrophobic (H) and hydrophilic (P) beads (residues). Hydrophobic residues attract each other, while the remaining possible pairwise interactions are equal to zero, except for excluded volume interactions. The model assumes that hydrophobic interactions play the predominant role in protein folding. This view was recently questioned by Honig and Cohen [91]. They argue that interactions involving backbone hydrogen bonding are also important. Scheraga and coworkers have also questioned the balance of entropy in these models between the native and nonnative states [92]. In related studies [38, 78, 79, 81, 93-106], the same interaction strength was assumed for HH and PP pairs, while interactions for HP pairs were somewhat weaker. The model has also been generalized to include all 20 amino acid types.

Using these models, a number of general issues were addressed that include the origin of the uniqueness of the native state [79, 89, 99]. For some sequences, the collapse transition was very cooperative, while for others, it was continuous [107]. The folding pathway(s) changed as well [99, 108]. Uniqueness is facilitated by incorporating a larger number of amino acids [79, 99, 104, 109]. Cubic lattice protein models have been used to search sequence space and to "design" fast folding optimal sequences [78, 106, 107]. In this context, a variety of reduced models that stress various aspects of the physical forces governing protein folding [108, 110-112] and dynamics [34, 40, 108, 113-117] were proposed. These investigations have provided interesting insights into the protein folding process and have motivated the idea of a folding funnel and the "new view" of protein folding [113, 114]. Sir are many excellent papers on this topic, we refer the reader to the literature.

The HP model assumes that protein folding is driven by lo interactions and that short-range conformational propensities are only in structural fine tuning [80]. Earlier, a different viewpoint emerg studies of simple diamond lattice systems and the chess knight model 67, 115-125]. These studies concluded that the native conformation from the interplay of secondary structural preferences and tertiary inte As Go et. Al. found in their very early work [65, 66, 126, 127], folding cooperative when the long- and short-range interactions are consistent native fold. Thus, while some features of protein folding ar independent, others such as the balance of interactions and confor entropy depend on the type of model used and the assumed forn conformational interactions.

1.4. Threading approaches to tertiary structure prediction

The early 1990's saw the development of threading technique attempt to assign a sequence to the best structural match in a library o solved protein structures [29, 128-140]. This approach is designed to proteins having little or no apparent sequence similarity to any of the in the structural library. Thus, it is designed to extend classical h modeling [141, 142]. A fundamental limitation of threading is that c have an example of the native topology already in the structural otherwise, the method cannot be fully successful. Threading might be of as finding the "least worst" match between the probe sequence library of template structures. Doubtless, its reliability will increas number of solved protein structures grows [143, 144]. In practice, to r problem computationally tractable, numerous simplified representatio protein have been developed, e.g., the molecule might be described by interacting Cαs or Cβs [129, 133]. In quite a number of cases, these have been successful at identifying the native topology. Often, tho native topology is not the best match but lies amongst the handfu scoring sequence structure matches [145]. Interestingly, even when tl topology is identified, the actual alignment of the probe sequenc template structure is often quite far from the best structural alignme probe and template proteins. This is an unsolved problem that is the s

intensive investigation [146, 147]. As a result, models whose backbone coordinate mean square deviation, cRMSD, from native that is on the order of 6-8 Å are typically generated. In other words, low-resolution models of the native structure are produced. Nevertheless, the advantage of threading approaches is they are quite rapid so that they can be applied on a genomic scale. In view of the excellent number of reviews available on the threading, we refer the reader to the literature for a more detailed discussion of this vibrant and important field [30, 148-153].

1.5. Exact restraint models of proteins

More recently, approaches designed to predict tertiary structure starting from random conformations and using a small amount of additional restraint information have been developed [49, 72, 154-159]. We term approaches that use experimental information "exact restraint models", a detailed discussion of explicit realizations of these models is presented in Section 2. Such exact restraints might include knowledge of secondary structure and/or some tertiary contacts. Such information could be provided from low-resolution X-ray crystal structures by NMR or by some biochemical means, e.g. the presence a disulfide cross-link. However, it is important to recognize that there are different types of "exact restraints". Quite often as a prediction exercise investigators assume that they know the secondary structure at the level of the exact angles [49, 67]. This can impart unrealistic expectations as to how the model will behave when restraints at the degree of resolution that can actually be provided by experiment are used. For example, one might know that a turn is present but not the chirality of the turn. In such a situation the native topology and its topological mirror image will be recovered (e.g., where the chirality of helices is right handed but that of the turns reversed), whereas if one assumes the chirality of the turns, no topological mirror images are generated. This might lead one to incorrectly conclude that such topological mirror images (which have a very similar pattern of side chain contacts, burial and secondary structure) are irrelevant. On the other hand, if one simply knew that there are three helices separated by two turns, then the problem of dismissing this alternative structure immediately emerges.

1.6. Restraint free *ab initio* protein folding

The most general approach attempts to predict protein tertiary structure from sequence without any recourse to known protein structures or evolutionary information. This is the traditional approach to the solution of the protein folding problem. We term this approach *restraint free ab initio* protei In its most pristine form, one uses the laws of physics to fold a prc scratch [27, 57]. However, due to its inherent difficulty, as a practic such models might also include some general knowledge based pote would not include any evolutionary information [28, 35, 45, 59, 63 161]. Examples of such knowledge-based contributions are an empiri energy scale [162] and knowledge-based pair potentials [163]. Obvic is the most difficult means of predicting protein structure, and as o Section 3.1, it has met with only quite limited success. When successf pristine approach could not only be applied to the problem of prote structure prediction, but also to prediction of the mechanism of prote and to provide insights in folding thermodynamics. Numerous pr applications of these ideas are described below.

1.7. Evolutionary-based approaches to protein structure predictio

Midway between the exact restraint models and restraint free protein folding models are what we term "evolutionary" based appr protein structure prediction, where no known structure of a protein ho to the protein of interest has been solved. This class of models uses s and tertiary restraints derived from multiple sequence alignments [84, Ortiz, 1998 #1069, 166, 167]. Such information might include secondary structure [168], tertiary contacts extracted from residue cor [164] or correlated mutations [169, 170], and knowledge-based pair derived from multiple sequence alignments [171]. Unlike the exact models, there are likely to be substantial errors in the restraint inform example, the average accuracy of predicted secondary structure is at with comparable prediction accuracy for tertiary restraints within ±3 [166, 172]. Thus, evolutionary based approaches have to accommoc and partially incorrect information; such models may require s modifications from their exact restraint counterparts. When exact resi used, then restraint violations can be used to eliminate a possible sti being relevant; when inexact restraints are employed, then restraint may occur in the entire set of distinct low energy topologies and canno for native topology selection [166, 172]. In Section 3.2, we di apparent physical origin of correlated mutations and describe the curr of the field of evolutionary based approaches to protein structure pred

2. EXACT RESTRAINT MODELS

2.1 Secondary and tertiary restraints in assembly of protein structures

As mentioned above, the present state of the art does not enable a dependable *ab initio* prediction of the majority of protein structures [173]. The most likely reasons are that the existing force fields are not specific enough and/or the sampling schemes are not sufficiently efficient at finding the native-like state of a model protein chains. Certainly, many recently designed interaction schemes for reduced protein models have captured a good part of the protein physics [28, 35, 39, 42, 43, 45, 59, 75, 85, 174-185]. This may lead to the conclusion that in a situation where the protein conformational space is reduced to its more relevant parts and when the sampling process is somehow guided towards the nativelike state, then the process of structure prediction will become much easier. This goal could be achieved by building into the protein model (and the sampling scheme) some secondary and tertiary restraints. These may come from sources such as CD and NMR spectra [186], protein crystallography, cross-linking experiments [187] and other experimental techniques. Approximate restraints of a similar sort could be also derived from various theoretical considerations. In this section, we focus on the simplest case, where the restraints (short-range - secondary and long-range - tertiary) are exact; however, they may be known at various levels of accuracy. Below, we discuss possible ways of implementing such restraints.

The meaning of the reduced models of protein structures is discussed in more detail later in this chapter; however, we stress here that the minimal requirements for a low-resolution structure prediction to be correct are: (A). The overall topology (the shape of the main chain trace) of the fold is the same as that seen in the experimental structure; (B). The obtained secondary structure is very close to that seen in the native structure and the alpha carbon trace root mean square deviation (RMSD) from the native structure is in the range of 3-7 Å depending on protein size. This level of accuracy may be of some use for application to protein function annotation [179].

Let us first discuss the short-range restraints, which could be described at different levels of accuracy. Suppose that the secondary structure is given by a three-letter code, i.e., helix, beta, loop, where "loop" stands for everything besides helical or β–type structures. Knowledge of the protein's secondary structure could be complete, or only some helical and β–type fragments could be known. Then, the remaining fragments of the polypeptide chain are treated as "loop" regions. Such a three-letter secondary structure code translated into structural restraints in many ways. One possib construct a set of potentials that drive the model chains towards a helix or beta strand. This could be done assuming the ideal target the ϕ, ψ angles. The target local geometry can also be made amino of amino acids is a better choice) specific. Moreover, the target local may involve all main chain atom types, only alpha carbon atoms, o side chain center of mass positions. The suitability of a particu depends on other model features. Local conformational biases superimposed on distances (and angles) involving pairs of residues, sequences and even longer fragments. An alternative (which could as a complementary set of restraints) method of implementation of structure target involves main chain hydrogen bonding. Known (o known) secondary structure translates into specific restraints for the r hydrogen bonding patterns. Helices should have a characteristic short-range hydrogen bonds; a helical residue within a helix cannot distance (along the chain) hydrogen bonds. Residues assigned to cannot form hydrogen bonds with residues assigned to be heli restrictions immensely suppress the conformational space of the mo that has to be searched.

Similarly, the long-range restraints could be implemented as potentials, square well potentials, or combinations of the two. T involve specific atoms (as alpha carbon atoms) or centers of mass of atoms (for instance that of side chains in particular rotameric state appropriate, the long-range restraints could be designed to closely spatial resolution of various NMR signals. The aforementioned cha protein hydrogen bond patterns could also serve as a framewo definition of long-range restraints.

To assemble a nativelike structure at a given resolution, a go model with an efficient force field should require fewer restraints th generic model that relies solely on the driving force derived from the Obviously, when the number of restraints is small, the resoluti obtained structures will depend mostly on the quality of the protein its force field. In contrast, when the number of restraints is very large model quality will depend on the restraint representation and th resolution of the model. Different classes of globular proteins of a s may require a different number of restraints to achieve models of c

quality. Helical proteins have much less conformational freedom, provided that the helices have been assembled, than do β–proteins of the same size. This suggests that in most applications, β–sheet proteins might require a larger number of long-range restraints. All these suppositions are strongly confirmed by the studies of different computational models of protein structure assembly outlined below.

2.2. Models with exact secondary structure but no tertiary restraints

One of the common ways of predicting protein tertiary structure assumes that the secondary structure has to be known before the prediction of a three-dimensional fold can be attempted [154-157]. While this view as a paradigm for protein structure prediction could be challenged, it certainly provides a straightforward framework that may sometimes prove to be useful. Indeed, there were a number of early attempts to apply such a methodology to low resolution protein fold predictions that were quite successful in some specific cases [154-157]. However, only recently has the problem of protein structure assembly, given its secondary structure, been more systematically addressed.

A very interesting method of protein fold prediction has been proposed by Dandeker and Argos [48, 49]. They consider all the backbone atoms and implicit side chains. The geometry of the main chain has been restricted to a small set of canonical values of the ϕ, ψ angles for various secondary structure motifs. A genetic algorithm has been used as the search method for the lowest energy state. In most of their computational experiments, the exact knowledge of the secondary structure (taken from known structures) and an idealized pattern of side chain hydrophobicity along the polypeptide contour has been assumed. Employing these assumptions, correct low-resolution structures have been successfully assembled for 19 small proteins that were representative of various structural classes. However, it has been observed that use of predicted (inexact) secondary structure led to a significant decrease in prediction accuracy.

Monge et al. [188] also assumed exact knowledge of the geometry of the regular secondary structure elements (helices in this case). Tertiary interactions have been modeled via a pairwise, knowledge-based potential for the Cβ–Cβ interactions. Then, the Monte Carlo method employed rotations within the loop regions to search conformational space. Their search process assembled only compact conformations. Moderate resolution (4-5 Å of RMSD from native) structures of four highly helical proteins have been obtained as the lowest energy structures. Subsequently, Gunn et al. [189] and Mo [177, 190] demonstrated that the exact knowledge of the short-range of regular secondary structure fragments allows for the low- assembly of more complex folding motifs. These included the α-helic myoglobin and the α/β fold of 66 residue C-terminal fragment c ribosomal protein.

A distance geometry approach could be quite efficient in proteir assembly [191] when the some of the secondary structure is exact Mumenthaler and Braun [164] attempted a test prediction on eig proteins, with the exact distance restraints within the helical fragme model assumed ideal helix geometry and a single (most probable representation of the side chains. The long distance restraints l applied in a very approximate way. First, they used multiple alignments for the statistical prediction of the buried and somewha residues. The average distances between pairs of buried, exp buried/exposed residues of a given type have been extracted from the database. Consequently, these restraints were rather inaccurate and filtered via a proper self-correcting distance geometry calculation. approximate character of these long-range restraints may be consic kind of long-range, mean force burial potential. Interestingly, ther some specific protein-like features encoded in these fuzzy lo correlations since in six test cases, structures with 2-3 Å RMSD fr (for helical fragments) have been correctly identified. The procedure two proteins with longer loops in spite of the large number o restraints.

Chelvanayagam et al. [159] also employed distance geometry. assumed known secondary structure and applied their approach to ei disulfide-rich β proteins. When the topology of the β–sheets, the ex of particular strands and the cross-link patterns were assumed kn their algorithm rapidly assembled the test structures by a proper filtra putative distance limits near the cross-links and within the β–sheets. β–sheet topology was assumed unknown, a combinatorial procedure a small number of possible native structure candidates.

A somewhat similar assumption of exact knowledge of secondary as employed in the aforementioned continuous space models, was al an early lattice Monte Carlo model due to Skolnick and Kolinski Godzik et al. [192]. However, there, the preferred local geometry wa

for the entire chain, thereby providing a weak bias toward the target, nativelike secondary structure. A Miyazawa-Jernigan hydrophobicity scale approximated the long-range interactions [163]. These studies demonstrated that even a small, but fully consistent with the native structure, secondary bias facilitates the very rapid structure assembly of plastocyanin and two TIM type α/β barrels. Similar to the work of Monge et al. [177, 188], further simulations of idealized folding motifs [125] showed that the target conformation for the loop/turn regions need not be specified.

All the work outlined in this subsection required exact knowledge of a substantial part of secondary structure for the successful assembly of three-dimensional structures. In all cases, where tested, the quality and/or reliability of these predictions substantially deteriorated when predicted (inexact) secondary structure data were assumed. This suggests that the tertiary interactions encoded in these models were not very specific. It appears that the models of tertiary interactions were good enough to select for the proper packing of well defined and relatively rigid secondary structure blocks; however, they have difficulties in correcting any substantial errors in secondary structure assignments. Indeed, in the latter case, the conformational space of the given model significantly increases, and the requirements for an interaction scheme are much greater. In this context, Monge et al. [177] proposed a very interesting way for evaluation and improvement of the tertiary interaction schemes for such reduced models. They reconstructed all atom structures from the predicted low-resolution folds and have shown that after a proper refinement process with a molecular mechanics potential within a continuum solvent approximation, it is possible to identify structures that are closer to the native one. An improved potential for the reduced model has been subsequently derived that has the form of van der Waals interactions between entire residues [190]. Whether this and other efforts to improve the tertiary contributions to the interaction scheme will enable structure prediction given inaccurate secondary structure remains to be established. At present, these works increase our understanding of the interactions stabilizing protein structures and controlling their assembly processes. Moreover, they may have important applications to protein structure determination from fragmentary experimental data.

2.3. Models with exact but loose secondary structure and restraints

Given a small number of distance restraints, several quite approaches to protein structure prediction have been recently publisl small number of restraints we mean that the number is small in c with that required for a standard distance geometry/molecular approach to the determination of protein three-dimensional struct NMR data [193].

Smith-Brown and co-workers [158] studied the folding of sev proteins given their native secondary structure and a number of l restraints. They used the Monte Carlo sampling method for an all-a chain model. The values of the main chain angles were kept near values for the given secondary structure fragments. Side cha neglected. The long-range restraints had the form of biharmonic between the Cα atoms. Due to the sequential implementation of the the long-range restraints, the simulation procedure assembled th structure in a specific order. First, the secondary structure formec extended conformation. Then, the long-range restraints were imp between a pair of adjacent elements of secondary structure to bi together. Subsequently, the remaining elements of secondary struc docked to the growing nucleus. The final stage of the simulatior possible distortions of the secondary structure geometry. The best structures had 3-5 Å RMSD from the native structure for the backbo Such results have been achieved when quite a large number of restra used. For instance, in the case of flavodoxin, 147 restraints were With a smaller (61) number of restraints, the flavodoxin structure dc 12 Å from native, yet still satisfied all the restraints. Using predicte of exact, secondary structure required an even larger number c restraints. These results suggest that these simulations were driv exclusively by the distance restraints.

Aszodi et al. [165], employed distance geometry and a polypeptide chain representation. Their model chain consiste Cα backbone and the Cβ positions of the side chains. They also knowledge of secondary structure and a limited number of exact l restraints. These have been supplemented by a set of "soft" somewhat in a similar spirit as those used in the work of Mument Braun [164]. They found that in order to assemble low-resolution s

more than N/4 exact tertiary restraints were necessary. Unfortunately, the algorithm generated not only structures of acceptable quality of about 5 Å RMSD from native, but also structures that have a 10 Å RMSD from the native. All satisfy the restraints. No clear method of selecting the proper fold has been presented. Subsequently, Aszodi and Taylor [194] applied this method in homology modeling. In this case, a large number of long-range restraints have been extracted from an alignment of the query sequence to a homologous sequence of known structure. The resulting restraints were weighted according to residue conservation criteria deduced from multiple sequence alignments. Models of quite good quality have been generated for several test proteins.

Bayley and coworkers [193] applied a combined genetic algorithm (GA) followed by simulated annealing to build molecular models from full NMR data for small proteins. A very large number of restraints were used, and the obtained structures were of very good quality. When the number of restraints was reduced (to ca. 10 per residue in case of BPTI), the majority of the GA calculations led to misfolded structures. Interestingly, the correctly folded structures (25% of all structures) were of similar quality as those generated for the full set of restraints.

Skolnick and coworkers [72] applied a high coordination lattice model for protein structure prediction from known secondary structure and a small number of known tertiary contacts between side chains. The protein representation assumes a Cα trace restricted to a lattice that allows 90 possible orientations of the virtual Cα-Cα bonds. The spatial resolution of this model is 1.22 Å, and the average cRMSD of the crystal structures fitted to this lattice is about 0.6-0.7 Å. The side chains are represented by a proper set of single sphere rotamers that mimic closely the rotameric spectra of real side groups. A knowledge-based force field has been developed for this model that enabled the *ab initio* computer folding of several small, topologically simple proteins [35, 59]. Since the force field of the model captured some essential features of protein interactions, it was expected that the model's applicability could be considerably expanded when a loosely defined secondary structure and a small number of long range restrains were used to guide the Monte Carlo simulated annealing process. Indeed, it has been shown that this MONSSTER algorithm (Modeling of New Structures from Secondary and Tertiary Restraints) enables efficient structure assembly given as few as N/7 to N/4 long range restraints for small globular proteins. A larger number of restraints (N/4) were required for β–proteins. In typical simulations, the resulting structures could l clustered, with two well-defined clusters of nativelike and topologi image structures and rare randomly misfolded structures. The structures had a backbone RMSD from native that ranges from 3-6 Å important is that the proper fold could always be identified from th conformational energy obtained from isothermal, low temperature M simulations of the folded (or misfolded) structures. Very similar r have been obtained with the same lattice model with a slightl interaction scheme (more explicit hydrogen bond representation an pair potential [45, 60, 85]) by Kolinski and Skolnick [85].

While the MONSSTER method seems to compare favorably witl related work, it still has some disadvantages. Firstly, the assembly proteins requires a larger number of tertiary restraints than for h proteins. Secondly, the cost of computation grows rapidly with pro length. Consequently, the simulation of proteins having more tha residues becomes computationally expensive when one takes into con that several simulations are needed to determine the proper structure fidelity. For these reasons, Kolinski and Skolnick [86] attempted to much simpler model of comparable accuracy. The model employ representation of a hypothetical chain connecting the centers of m side chains in their actual rotameric state. The underlying cubic lat spacing of 1.45 Å, and the length of chain segments varies from 4.8 It covers the main portion of the distribution of distances between adj chains in real proteins. A cluster of points on the cubic lattice repre residue, and which allows for the very rapid and straightforward h the model's excluded volume and detection of side chain contacts. T has built-in knowledge-based potentials for short-range conf propensities, a one-body hydrophobic potential, pairwise potential an of generic (sequence independent) cooperative interactions th hydrogen bonds. The force field is good enough to drive *ab* resolution folding of very simple, small globular proteins. The structure is encoded by weak biases in short-range potentials for the l β–fragments, and proper selection rules for main chain hydrogen bon a β–residue cannot be hydrogen bonded to a helical residue). The l restraints have the form of harmonic potentials. Additionally, the pairs of residues have interaction parameters that are modified (by constant) with respect to the original statistical pair potential. D

simplicity of protein representation and the form of the force field, Monte Carlo sampling is very fast and scales favorably with the chain length.

Using this algorithm, all types of proteins could be folded with the assistance of N/7 long-range restraints, where N is the number of amino acids in the polypeptide chain. For example, the 56-residue protein G could be folded with 8 restraints to structures that cluster around 3 Å cRMSD from the native. For the sake of easy comparison with the previous work, the cRMSD is measured for an approximately reconstructed Cα trace. It is worthwhile to mention that the cRMSD for the side chains is only slightly worse. Some structures could be reproducibly folded with an even smaller number of restraints (4 in the case of protein G, while 2 restraints are sufficient only when properly selected). This model allows for the assembly of much larger structures and has been tested for proteins containing up to 247 residues. The accuracy of the assembled structures depends on chain length and the number of tertiary restraints. For the 68-residue 1ctf fragment (10 restraints), the average structure has a 3.2 Å cRMSD from native; for a 108-residue 2trx fragment, the cRMSD is 3.1 Å, with 30 restraints and 3.5 Å with 16 restraints. Similarly, for a 138-residue 3fxn fragment (20 restraints), the cRMSD is 4.1 Å and for the 146-residue 1mba protein (20 restraints), the cRMSD is 4.3 Å. Finally, for the 247-residue Atim, the backbone cRMSD ranged from 5.1 to 6.7 Å, depending on the number of tertiary restraints. The restraints were generated randomly (however, very close contacts along the chain were rejected). In many cases, some adjacent pairs of secondary structure elements did not have even a single long-range restraint. These results were possible due to the important contribution from the model interaction scheme. Indeed, when only the restraints and the secondary structure biases were used, the results from this model are much worse and comparable to other studies [165]. Some test (unpublished) simulations show that when the exact secondary structure is replaced by the predicted ones, the resulting structures are in most cases essentially the same; however, β–proteins in the limit of a low number of tertiary restraints produce slightly worse structures, and the fraction of misfolded structures increases. In all cases, the proper fold can be identified based on reproducibility and the average conformational energy extracted from low temperature, isothermal Monte Carlo trajectories. Two representative examples of the structures generated with the smallest number of tertiary restraints are compared with PDB structures in Figure 1.

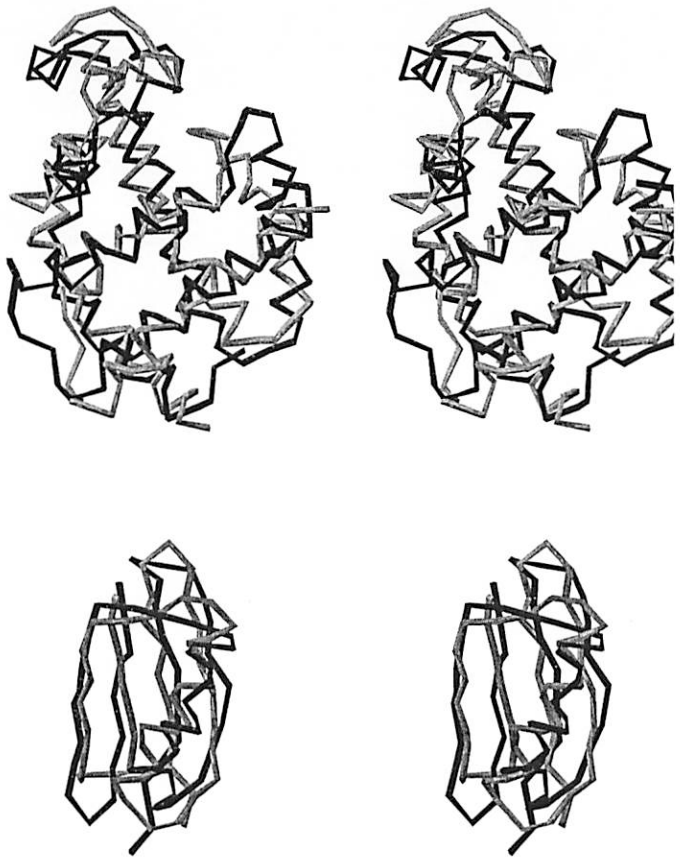

Figure 1 Stereo drawings of 1mba and 1gb1 in upper and lower p respectively. The black lines correspond to alpha carbon trace ex from PDB files. The models obtained in MC folding simulatio shown in gray.

2.4. How can these approaches be integrated with experiment

The outlined methods for assembly of protein models have important applications to structure determination from sparse experimental data. The most is to model building from rather complete NMR data. Work by Bayley and coworkers [193] is a good example of a novel approach to this well-defined and standard problem.

For some cases, it is easier to experimentally determine the secondary structure of a native protein without any information about tertiary interactions. The work described in Section 2.2 is aimed at structure determination from just knowledge of secondary structure [48, 49, 159, 164, 177, 188-190]. For regular helical proteins, these methods may provide low-resolution models. For more complex topologies, the probability of building a good molecular model just from secondary structure information is rather small.

Perhaps more interesting are those models that allow for structure assembly from known secondary structure and sparse experimental data on tertiary restraints [72, 158, 165]. Such situations may arise in the early stages of protein structure determination via NMR techniques. Some algorithms [86] described above can build low-to-moderate resolution models from as few as N/7 (or less in simpler cases) tertiary restraints, where N is the number of amino acids in the protein. Such a model could be further refined by more detailed computations. On the other hand, the model provides quite accurate predictions for all other possible tertiary interactions. Such predictions may be useful in the process of deconvoluting NMR spectra and may suggest directions for further experiments that are aimed at structure rectification. The process may be iterated. Sparse experimental data lead to an approximate model. Then, the model suggests how to assign (or validate via another experiment) additional tertiary restraints. The larger set of restraints enables the assembly of a more accurate model, etc. It is worthwhile noting that various types of tertiary restraints, which correspond to various NOE signals, could be easily encoded in most of the reduced models described in the previous section. Preliminary work based on the Kolinski and Skolnick model [86] shows that encoding not only signals involving side groups but also other signals increases the quality of the obtained models for a given number of tertiary restraints. For example the NOE's between alpha carbon hydrogen and amide hydrogen encode not only contact information but also quite rich directional information. This significantly improves the quality of β and α/β protein models.

Another interesting possibility opened by these models allows for re
structure assembly based on the predicted secondary structure and a f
range restraints may be disulfide bonds. Having an approximate mc
may suggest a limited set of plausible point mutations to introduce a
cross-links. This way, the probability of determination of a side cha
chain contact in the protein of interest in a single experiment is
increased in comparison with more random cross-linking experiments.

Future applications may involve chain tracking and model building l
low resolution X-ray and electron microscopy data. Suppose that an
density map allows one only to assign with a certain level of confid
possible positions of some large amino acids with many possible ma
pathways that satisfy such low-resolution data. Such a restraint set
sufficient for structure assembly, provided the model force field
enough to guide sampling of plausible proteinlike conformations.

3. TERTIARY STRUCTURE PREDICTIONS BY *AB INITIO* M BUILDING

3.1. Predictions by restraint free folding.

Tertiary structure prediction by restraint-free simulation has constitu
many years, the *classical* approach to the Holy Grail of obtaining
structure from sequence information [27]. Because of its intellectual ap
enormous potential implications, it has been, and still is, an area o
research. However, the advances produced by this approach have bee
modest. The reason is that a full solution to the two convoluted prob
sampling conformational space and generating an energy landscape v
global minimum at the native conformation needs to be obtained for
Despite some progress, such a formidable accomplishment appear
unreachable nowadays for all but some of the most simple protein mo
35, 59, 83, 161, 178, 195]. Still, research in this area is required bec
one hand, the resulting theoretical models constitute first approximatio
which other more pragmatic structure prediction techniques that inc
knowledge-based information or additional heuristics can be built.
other hand, the lessons obtained from them can teach us general pr
about the kinetics and thermodynamics of protein folding and their rela
with sequence and structure that can be of utility in other areas, such as
engineering [196].

The different research groups, using different levels of detail have employed a range of different models. Models that have only one interaction center per one, or even per two, residues provide extreme simplification [28, 50, 65, 66, 75-81, 197]. At present, it is unclear whether these models possess sufficient geometric resolution so that an effective energy function can be constructed based on them. Nevertheless, they are useful for fast model building when additional information is available in the form of distance restraints or more generally, for exploring general aspects of folding kinetics and thermodynamics. At the next level of complexity, a popular medium resolution model that achieves substantial reduction in complexity yet retains substantial geometrical fidelity is based on collapsing the side chain atoms into a unique pseudoatom, usually located at the center of mass of the corresponding side chain [35, 69, 160]. Backbone atoms, on the other hand, can be used in full atomic detail or they can also be collapsed into a unique pseudoatom located at the Cα position [85]. In this last case, pseudodihedral angles connecting the backbone pseudoatoms are used as internal coordinates. More detailed models, used by some groups, are based on an atomic level description of the protein coupled to continuum models for the description of the solvent [198]. Very recently, there have been also some impressive advances in the use of full atomistic description models, including explicit solvent [57, 199]. In what follows, we give a short review of the latest advances using each one of the models for globular proteins.

Highly simplified models of protein structure embedded into low coordination lattices have been used for tertiary structure prediction for almost 20 years [65, 66, 75]. For example, Covell and Jernigan [64] enumerated all possible conformations of five small proteins restricted to fcc and bcc lattices. They found that the nativelike conformation always has an energy within 2% of the lowest energy. Virtually simultaneously, Hinds and Levitt [28] used a diamond lattice model where a single lattice unit represents several residues. While such a representation cannot reproduce the geometric details of helices or β-sheets, the topology of native folds could be recovered with moderate accuracy.

The pioneering work in the use of medium resolution protein models to predict protein structure *ab initio* is due to Levitt and Warshel [75]. They were able to obtain models of BPTI with a cRMSD from native of about 6.5 Å. The significance of these predictions was later called into question by Hagler and Honig [200], who obtained comparable quality structures using a glycine and alanine heteropolymer model of the BPTI sequence. Some time late and Doniach [160] developed a similar model that, when applied to yielded low-resolution structures, with several proteinlike features. Th of a number of other single domain proteins has also been examinec 56- and 90-neighbor lattices using medium resolution models deve Skolnick and Kolinski [35, 59, 63, 179]. Folding simulations of the I of protein A [35, 59] yielded structures whose cRMSD from nati ordered parts of the molecule is 3.3 Å. The folding of crambin, assuming the identity of the disulfide crosslinks, produced lov conformations having an average Cα cRMSD of about 4.0 Å.

High-resolution models have also been employed by a number of but most have involved small peptides. Thus, Pedersen & Moult have the application of genetic algorithms to the determination of protein from sequence using a full atom representation of the solute and a c model for the solute-solvent interaction [201]. Peptide fragments of residues long were tested, and it was found that, in most cases, the lo energy structures produced by the GA were similar in conformatic corresponding experimental structure. Avbelj & Fele using the "ele screening model" have attempted larger structures [202]. In their appr free energy of the protein contains two contributions: burial for all a side chain dependent electrostatic interactions for the backbone a "hierarchic condensation" algorithm is used based on Monte Carlo sin in torsional angle space, increasing the range of interactions during th of the simulation. The method can predict the local secondary structur as some supersecondary structure motifs and some small helical prot the other hand, some encouraging studies using a full atomic descripti solvent have also been recently reported. Daura *et al.* have descr reversible folding in solution of an heptapeptide by using molecular c simulations [199]. But so far, Duan & Kollman give the most in achievement in structure prediction using detailed models with explici [57]. They have been able to fold, using molecular dynamics simulat starting from a random extended conformation immersed in a box molecules, the villin headpiece subdomain, a 36 residue peptide, to a having a cRMSD with respect to the experimental structure of 4.5 result is most impressive considering that the simulation contains on t of 10000 atoms and simulated 1μs of the folding reaction, a time very experiment. Such impressive results could be obtained by a

parallelization of the MD code and a no less impressive use of a massively parallel Cray T3E with 256 CPU's during 2 months of computation.

3.2. Prediction by restraint driven folding: Evolutionary based approaches.

Approaches to structure prediction using different flavors of secondary structure constraints have been recently reported. This renewed interest in the use of secondary structure information as a way of reducing conformational space when building molecular models is the result of the recent improvement in accuracy of secondary structure prediction methods due to the introduction of evolutionary information.

A common theme successfully used by some of the authors is the incorporation, in some way or another, of supersecondary structure constraints. This is an interesting strategy, as it reduces the conformational space by eliminating many non protein-like states, and allows the sampling to be focused on the most relevant regions. For example, Cui et al. have published a method that assembles, using a genetic algorithm, predicted secondary structure elements using a library of clustered supersecondary structure motifs [203]. A second innovation in their method is the application of a fast algorithm for the computation of the approximate accessible surface area of the conformation. Thus, a "physical-based" force field is applied that also includes hydrogen bonds. Tests using several small proteins showed that native like conformations can be assembled as the lowest energy states. Along similar lines, Jones has described a method that has been able to successfully predict blindly the fold of the NK-lysin during the CASP2 contest [204]. The method is based on the assembly of recognized supersecondary structural fragments by the application of a Monte Carlo based simulated annealing algorithm. The force field in this case is knowledge-based and is extracted from a statistical analysis of the protein database. A slightly modified approach, but similar in spirit to that of Jones, has been reported by Simons *et al* [205]. Their method consists of assembling structures from a library of fragments extracted from a structural database. The fragments are extracted on the basis of a profile based similarity measure of the local sequence, and consist of secondary and short supersecondary structure motifs. A Monte Carlo method using a Bayesian scoring function extracted from the protein database is used to assemble the structures. The method has been able to assemble some complicated motifs, particularly in the case of helical proteins. However, as been noticed others, selection of the native like topology cannot be reliably done.

A different approximation has been introduced by Samudrala et a Their method blends a combination of approaches, including s structure predictions, in an interesting fashion. First, using a tetrahedr all possible self-avoiding conformations of small proteins are exh enumerated. The best scoring 10000 conformations are selected using based scoring function. Then, for each conformation, using idealized l sheet values based on the predicted secondary structure, all atom struc generated by fitting an off-lattice four-state ϕ/ψ model to th conformation. The all atom conformations are energy minimized and e using a second hybrid scoring energy function. The best scoring m used to generate consensus interresidue distances that are used in a geometry algorithm to generate the final predicted conformation. The has been tested on a set of 12 small proteins, giving good results fo proteins for which accurate secondary structure predictions were availa

3.2.1. Evolutionary approaches based on residue conservation

Secondary structure prediction methods have recently shown con improvement when many evolutionary related sequences are avail analysis [206, 207]. It is natural to ask whether additional informati the arrangement of the secondary structure elements in space can be from the analysis of the multiple sequence alignment. This idea then leads to trying to predict tertiary contacts from the alignme fundamentally different points of view have been applied to this probl first is based on the analysis of residue conservation and the second is the analysis of residue covariation. The study of conservation ha history in sequence analysis and has been applied in different context family classification or to binding site identification [20]. Several auth also applied this idea to the prediction of tertiary contacts [164, 169, 1 basic idea is that totally conserved hydrophobic residues should important role in determining protein structure, and most probab residues are found in the protein core. Thus, one could in principle esti contact distance between two residues under the assumption that the form part of the protein core. Such distances can be calibrated as a fu the protein size using a database of known proteins and sequence align

Contact prediction by analysis of sequence conservation has been used by Aszodi & Taylor [165] and also by Braun and coworkers [164]; both were previously discussed above in more detail in the context of studies that employ exact secondary structure assignments. In both cases, the set of restraints are used in distance geometry based algorithms (metric matrix based in the case of Aszodi & Taylor and target function optimization in dihedral angle space in the case of Braun and coworkers) in order to assemble the protein fold. The problem with this type of approach is the lack of structural specificity in the contact prediction; i.e., a large contact distance needs to be used in order to avoid a large number of false positives. Thus, contact prediction on the basis of conservation can only be used either as a general regularizer (the Aszodi & Taylor approach [165]), or additional techniques during the simulations are required to eliminate the false positives if smaller radii are used. For example, Braun and coworkers [164] have suggested the use of what they call the self-correcting distance geometry method, in which subsets of restraints are left out and the impact on the force field energy tested. Restraints for which significant improvement in the energy function is observed are left out. This approximation assigns a very important role to the protein force field in the evaluation of the restraint quality. It is not clear whether with current energy functions, sampling schemes and errors in the predicted restraints, this is a feasible strategy. As a result, the best results using residue conservation are obtained with small helical proteins.

3.2.2. Evolutionary approaches based on correlated residue mutations

The study of variation in the multiple sequence alignment could in principle provide a more sensitive method for the prediction of specific contacts than the study of conservation. A technique to perform such a difficult task is to look for correlated mutations. The underlying idea being that a significant correlation above the background in the alignment may imply closeness in space for the correlated positions. However, the prediction of contacts from correlated mutations in protein structures is a controversial subject. Several studies have been made, with evident discrepancies in their conclusions. Jones [208], in his review of the CASP2 contest in 1997, states that "the consensus opinion today is that, whilst it is certainly possible to predict specific contacts in protein structures from multiple aligned sequences, it is difficult to make use of this information because of the relatively large numbers of false positives that are thrown up". Rost & Sander [209] give a similar conclusion in their 1996 review: "So far, none of the methods appears to find a path betv Scylla of missing too many true contacts and the Charybdis of predi many false contacts". An analysis of the methods available indicated effects were mainly responsible for the poor performance of the cont prediction: the clustering of sequences in subfamilies (the "subfamily and the presence of indirect, or multiple, correlations among different p in the alignment.

However, the scenario is different today. Ortiz *et al.* have been develop a new method that overcomes these two problems and appea precise enough in the prediction of a small subset of contacts, which t comprise about 20 % of the entire contact map [84, 166, 167, 172, 210 contacts are, however, not accurate, and only about 70% of them are when an error of ±3 residues in the local vicinity of the real contact is But when they are correctly used as restraints in specialized folding alg they are of sufficient quality to fold small proteins to low resolution st in a significant fraction of cases. The method is based on a combin multivariate statistical analysis and local threading. The algorithm w two steps: First, a few tertiary contacts (termed *seeds*), between the se structure elements are predicted from the multiple sequence alignme program sequentially applies partial correlation (in order to eliminate effects) and factor analysis (in order to eliminate the subfamily effect Pearson correlation matrix derived from the alignment. Typically, for up to about 100 residues, about 5 seed contacts are selected. Next, the contacts are then *enriched* by threading the test sequence through a st database and then selecting pairs of secondary structure elements pred be in contact on the basis of the seeds. Then, energy and cRMSD cut applied to the selected fragments. If the set of fragments survives the procedure, then additional contacts found in the selected fragme projected onto the target sequence.

3.2.3. Incorporation of evolutionary information in MONSSTER

In MONSSTER, a secondary structure bias is incorporated using a function; for those residues having a predicted secondary structur energetic biases for the various allowed conformational states are [211]. Regions predicted as U-turns are assumed to lie at the protein [212]. For these residues, a penalty is added when they lie at or be radius of gyration. This term acts to reduce kinetic traps by segrega

different parts of the protein into its corresponding layers. The hydrogen bond potential is also modified for those residues assigned to a predicted type of secondary structure so that the resulting hydrogen bond pattern is compatible with the secondary structural prediction. In addition, a cooperativity term that stabilizes and propagates the formation of β-sheets is included in the potential. Incorporation of predicted contacts as restraints also demands a slightly different implementation from that used in the case of exact restraints. It is necessary to take into account the spatial resolution of the restraints, and the possibility of assigning wrong pairs of contacts. Thus, the restraint function consists of a simple flat-bottom harmonic potential, operating either between side-chain centers of mass or between the projection of the residue pair onto the principal axes of their respective secondary structural elements, an implementation termed restraint splinning. This implementation is often needed as a result of shifts in registration in the contact map prediction: most predicted seeds are shifted by at least one residue with respect to the experimentally observed contact, and after *growth,* the different patches of contacts can have different phases.

Furthermore, in order to have a better sampling of proteinlike regions, it is convenient to introduce knowledge-based restraints designed to reproduce the packing of supersecondary structural elements. This knowledge-based information acts to reduce the number of misfolded structures. Two types of knowledge-based rules are considered in MONSSTER, namely the chirality of βαβ units and the angle formed in ββα supersecondary structure units. However, if the number of loop residues is greater than 15 residues, it is assumed that the secondary structure prediction algorithm has missed an intervening secondary structure element, and the knowledge-based rules are not applied at all.

In order to obtain enough statistics, a large number of simulations is required. Normally, a series of up to 1000 independent simulated annealing runs are performed. Low energy structures are selected, typically the lowest 1% set of the complete pool of structures, and the resulting structures are clustered on the basis of their pairwise cRMSD. From these, we select representative structures from each of the families obtained, and proceed to the *native structure selection* stage which consists of isothermal runs from which the putative native topology is chosen on the basis that it has the lowest average energy.

Table 1.
Results of the structural alignments of some predicted structures with th experimental conformation using the structure superimposition program [213]

PROT	HIT[a]	Z_{sc}	RMS[c]	LA[d]	STRUCTURAL ALIGNMENT[e]
1c5a	2abk	2.0	3.0	48	3-8,12-15,20-23,28-38,42-48,49-54,56-65
					3-8,10-13,23-26,28-38,42-48,50-55,56-65
1cis	2sec-I	1.0	2.6	40	2-7,8-11,14-19,27-32,36-39,45-48,49-53,56-60
					1-6,9-12,13-18,28-33,35-38,44-47,49-53,60-64
1ego	1grx	4.3	3.0	68	1-7,10-26,27-38,40-45,60-65,66-85
					2-8,10-26,29-40,48-53,58-63,66-85
1fas	3ebx	0.4	3.8	41	5-8,17-28,30-37,40-56
					6-9,17-28,32-39,41-57
1ftz	1lfb	0.9	3.2	45	3-6,7-15,16-19,22-37,42-45,47-50,52-55
					3-6,8-16,19-22,23-38,42-45,46-49,50-53
1gpt	1sco	0.2	2.9	33	6-12,16-27,28-36,38-42
					8-14,16-27,29-37,39-43
1hmd	1cei (2hmr)	5.0	4.1	76	1-18,22-28,29-45,46-67,72-83
					4-21,22-28,30-46,49-70,72-83
1ife	1tig (1ife)	2.8	3.7	69	1-20,23-28,31-43,45-63,75-80,83-87
					2-21,23-28,30-42,49-67,71-76,83-87
1ixa	1edm	0.6	3.0	30	6-11,13-22,24-37
					6-11,14-23,24-37
1poh	1rth-A	1.9	3.2	57	2-5,15-29,31-34,42-45,53-56,58-61,63-69,70-80,82-85
					1-4,15-29,33-36,48-51,55-58,59-62,63-69,71-81,82-85
1pou	1oct-C	5.0	2.7	65	1-24,26-40,42-46,50-59,61-71
					2-25,26-40,41-45,51-60,61-71
1shg	1abo-A	0.9	4.0	43	4-7,8-16,25-28,29-32,36-57
					5-8,10-18,25-28,30-33,36-57
1ubq	1ubq	2.5	3.4	58	1-7,9-18,21-37,41-46,56-59,61-64,67-76
					1-7,11-20,21-37,39-44,45-48,59-62,66-75
3icb	1wde	3.7	3.3	64	1-19,25-31,32-36,38-58,64-75
					1-19,25-31,33-37,38-58,64-75
PrtA	1edl	2.7	2.5	42	2-15,17-23
					2-15,17-23

[a]First hit (according to the Z-score value) of the structural alignment of the predicted conformation a set of DALI representative folds of the protein data base. Bracketed names correspond to second hits
[b] Z-score of the structural alignment of the predicted and experimental structures, as defined in the D. method.
[c] CRMSD of the predicted and experimental structure for the aligned region.
[d] Number of residues used in the structural alignment
[e] Regions aligned. The residue numbering scheme refers to the sequential numbering from N to C ter to the actual PDB numbering scheme.

Using a test set of 19 small proteins, we have demonstrated tl approach can assemble native like topologies. The average cRMSD lowest average energy structures corresponding to the native topology from about 3 Å for some helical proteins to roughly 6 Å for β and α/β p The relatively high cRMSD between the experimental and predicte arises from registration shifts of the secondary structure elements cre errors in predicted restraints, poor positioning of the loop regions, ar regions where no restraints are predicted. Nativelike conformations

obtained either as the best average energy in 16 of the 19 cases studied or as the next best energy structure in the remaining three cases. However, in most cases, the standard deviation of the energy in a given structure is of the order of the energy difference between the average energy values, i.e., the energy spectra substantially overlap. Thus, selection of the native fold on the basis of the force field energy is uncertain.

Results of the structural comparison between predicted and experimental structures are presented in Table 1. For 14 of the 19 predicted structures, it is possible to find a structural alignment covering about 80% of the residues of the protein with a cRMSD of about 2 to 4 Å from native, but the residue fragments are shifted in registration between the predicted and experimental structures. Thus, we conclude that this *ab initio* folding approach produces structures of comparable quality to threading methods.

3.2.4. Physical basis of correlated mutations

Recently, Ortiz *et al.* have tried to establish whether there is some analogy between their approach to protein structure prediction and current knowledge of protein folding kinetics and thermodynamics [214]. By analyzing recent results from Shakhnovich's group on fast and slow folding model proteins (48-mers on a cubic lattice) [215], it was shown that correlated mutations at neighboring positions in three dimensional space (i.e., contacting residues), naturally arise as a consequence of the evolutionary pressure on proteins to rapidly fold to their global energy minimum. This conclusion is based on the observation that these correlations occur in positions that are close (in space) to the thermodynamically characterized folding nucleus of the model protein. It has been rigorously shown that a subset of the residues forming this folding nucleus is able to discriminate between fast and slow folding sequences, or in other words, it is responsible for the differences in folding rate of the sequences. A possible physical explanation of this effect is that these correlated mutations arise from an attempt by the system to minimize its frustration as it evolves in sequence space.

Predictions from these model studies match well with results on experimental protein folding studies of some real proteins (Ortiz and Skolnick, unpublished). Indeed, when a similar procedure is employed on real proteins for which experimental data are available, there is a substantial overlap between the folding nucleus found experimentally and the folding nucleus predicted from the multivariate analysis of multiple sequence alignments. Thus, once the seeds are expanded, we speculate that a substantial pr of the real folding nucleus of the protein is used as a restraint MONSSTER simulations.

Some other parallels between the current protocol to protein s prediction and our knowledge of folding of real proteins are worth not example, we have observed that in the folding simulations, a higher correct topologies is obtained when residues predicted as loops are "pul from the structural core to the surface with a biasing potential. It is of that in the analysis of fast folding model proteins one of the main responsible for the higher folding rate is what we have called a "loop ef which residues in certain loop positions are different in fast and slow sequences. Another interesting parallel is related to the number of r required for successful fold assembly. We have noticed that about N/4 are required to succeed in folding, and that this number can be obtain the expansion of seeds, whose number is about 5 for a 100-residue prot of interest that similar numbers have been observed in theoretical st protein folding. For example, recent lattice and molecular d simulations indicate that the number of contacts in the folding transitior of the order of N/4, and that the average number of contacts in the nucleus for small model proteins is 5. Given that we have demonstrate least a fraction of the predicted contacts by correlated mutations are adj the protein folding nucleus, it is tempting to speculate that part of the nucleus is included as a restraint during the simulations. Once the nucleus is arranged in space, the search for the native state is esser downhill process on the energy landscape. Thus, relatively simple forc should be sufficient in order to allow for the on-site construction of the the structure around the folding nucleus. This fact could explain the su structure prediction for small proteins of this method, and why only small number of restraints extracted from a very limited set of predicted using evolutionary information it is possible to assemble nat conformations.

All these findings rationalize the results obtained so far with the approach to restraint-based protein fold prediction, and link theoreti experimental studies of protein folding with theoretical approaches to s prediction. This is quite exciting, as the convergent points of view of t and experimentalists are beginning to have an impact on practical app to predict protein structure.

3.3. Limitations and outlook

While some advances in the field of structure predictions by *ab initio* model building are apparent, we are still far from having reliable methods for structure prediction, even at the level of small proteins. Useful checkpoints of the state of the art in the field are the CASP meetings, held regularly, at intervals of one or two years. These serve to evaluate in a large-scale experiment the accuracy of structure prediction methods. In CASP2, the most recent meeting in which evaluation data are available, ten groups participated in the *ab initio* prediction category [216]. In general, although there were some *ab initio* predictions that were reasonably close to the native structure, the results were disappointing. The best of these predictions came from Jones's laboratory, that was able to predict by Monte Carlo simulations using predefined building blocks of supersecondary structure elements and empirical potentials the structure of NK-lysin to a cRMSD of 6.2 Å [204]. Since then, the advances discussed in this review suggest that today the situation is more optimistic. Thus, it has been possible to make successful blind predictions of several small proteins. One well-documented example is the prediction by Ortiz and coworkers of the 81-residue KIX domain of the CREB binding protein. Contact map prediction followed by fold assembly simulations yielded either a left- or right-handed three-helix bundles. For the correct topology, a cRMSD of 5.5 Å with the experimental fold is obtained [217].

The CASP3 contest is now in progress, and results for some proteins are already available. From the partial results, it seems that substantially better predictions are possible than were done in the past, but we must wait until a full analysis of the results is available to assess the final outcome of this contest.

4. WHAT IS THE REQUISITE RESOLUTION OF PREDICTED STRUCTURES?

A key question in the field of protein structure prediction is how close must a given model be to the native state in order for it to provide useful information. In the previous section, we have shown that small proteins can in a fraction of cases be predicted with a backbone cRMSD of 4-6 Å. These are typical of the average cRMSD of threading models for larger structures as well. Such models have the same global topology as the native structure, but there are errors in chain registration and packing angles. Nevertheless, we argue that for many biologically relevant questions, such models are quite useful. To identify binding regions, one is interested in which residues are expos to identify binding epitopes), and here the accuracy is acceptable. Furth in at least a number of test cases, Skolnick and coworkers have sho these models can be used to identify the active site residues associated given class of chemical reactions (e.g. disulfide oxidoreductase activi 23]. On the other hand, because the interiors of these model proteins ar poorly packed and there are substantial errors in the side chain positio cannot be used to identify ligands. Such models are not appropriate fo lead compound identification. In other words, given the current state of models can be produced that have significant use in biology, and whic be used as initial structures for rapid NMR refinement, but cannot be u chemistry. A key question which must be answered is what resolution o is required so that lead compounds can be identified using conten approaches, not necessarily as being best, but within a reasonable th that can be used for screening. Alternatively, different molecular des could be developed that could be used with lower resolution models.

5. TECHNIQUES FOR LOW TO HIGH RESOLUTION MODEL

As indicated in Section 1, a possible approach to the solution of the folding problem is to use a hierarchical approach [63, 83, 84]. One starts reduced protein model and then assembles the overall topology. Then, detail is added. While quite reasonable in principle, in practice there a very few examples of success. Early work yielded mixed results. For domain of protein A, which adopts a three-helix bundle, the backbone c from native of the detailed atomic model did not show improvement fr initial reduced model [63]. However, for the GCN4 leucine zipper (a coil), the detailed atomic model showed substantial improvement, started from about a 3.7 Å cRMSD, the resulting backbone cRMSD of th structure was 0.8 Å [83]. However, its native conformation is very sim consists of the side by side association of two α helices. More re Simmerling and coworkers started from a model of a 29-residue p CMTI-1 generated by MONSSTER whose initial backbone cRMSD native is 3.7 Å [84]. Using the Locally Enhanced Sampling method cor with the Particle Mesh Ewald technique [218], they produced a structure cRMSD is only 2.5 Å from native. Of course, this is a very small prote more detailed studies on other systems must be done to establish the ger of this result. Nevertheless, to put these results in proper perspective, w

that models of this quality were commonly produced in the early days of protein NMR spectroscopy. Thus, while improvements and further validation are clearly necessary, encouraging progress is being made.

Quite often the structures generated from threading have insertions and deletions, (especially in loops which may be involved in binding). In fact, the alignments substantially differ from the best models that could be produced on the basis of structural alignments. Thus, Kolinski and Skolnick and coworkers have developed an approach that may allow for the refinement of models produced by threading [219]. The structure is refined in the context of a side chain based lattice model that employs a number of short and long range potentials derived from multiple sequence alignments. The starting conformation of the lattice chain approximately follows the aligned template fragments. Then, Monte Carlo simulated annealing is used to minimize a combination of the system's internal energy (as defined by the model force field) and the distance from a loosely defined tube surrounding the aligned part of the template chain. As shown in Table 2, for a number of test cases, after the models are minimized, there is considerable improvement in the quality of the model. Because it is reasonably rapid, requiring about a CPU day per sequence, see Section 6, it is applicable to whole genomes and nicely complements classical homology modeling techniques. Since this technique is very much in this spirit, we term it generalized homology modeling.

Table 2.
Results of refinements of the threading alignment based models by Monte Carlo simulations an a reduced model

Protein PDB code	Sequence length[a]	Full model cRMSD[b]	Threading alignment cRMSD[c]	Final alignment cRMSD[c]	Alignment length
1hom_	68	3.76	5.59	3.53	45
1tlk_	103	4.64	7.88	4.57	84
256b_	106	3.88	4.55	3.90	104
2azaA	129	9.40	11.04	10.45	80
2pcy_	99	4.37	7.76	4.43	93
2sarA	96	7.72	8.28	6.95	72
3cd4	97	5.96	5.72	5.49	79

[a]For 1hom, residues 8-60 are considered to be structured, for 1tlk residues 9-103 are considered to be structure; otherwise, the entire protein is compared in the Table. All RMSD values are for alpha carbon atoms.
[b]cRMSD from experimental structure after Monte Carlo refinement/model building for entire molecule (except for the unstructured parts of 1hom and 1tlk).
[c]cRMSD from native target structure of the threading-aligned fragments before and after Monte Carlo refinement. The last column gives the total length of the threading alignments (number of aligned residues).

6. ROLE OF STRUCTURE PREDICTION IN THE GENC REVOLUTION

The computational requirements for evolutionary-based foldir threading approaches to genomic scale structure prediction are substantial, but not unreasonable given the increasing availability of fa cost PCs. For example, contemporary evolutionary based protein methods are applicable to single domain proteins, up to about 15(residues in length and can identify possible novel protein folds [166, 16 Threading is significantly less expensive [132], but often there are in and deletions in the subsequent alignments that require sub modification using generalized homology modeling tools such as th have developed [219]. Table 3 gives a summary of the CPU requirem protein structure prediction on the genomic scale.

Table 3
Computational requirements in CPU days for protein structure predicti genomic scale

Genome	Number of ORFS	Number of ORFS <150 residues	Monsster Folding CPU time[a]	Threading CPU time[b]	Refinem time[c]
M. genitalium	408	82	4,920	2	4
H. Influenzae	1,680	369	22,410	8.4	1,
M. Jannaschii	1,735	425	25,500	8.9	4
E. coli	4,290	879	52,740	21.5	4,
S. cerevisiae	18,567	1433	85,980	92.8	18

[a] Assumes 1000 folding simulations with an average 60 CPU days per sequence on a single proces SGI ORIGIN 200 running at 180 megahertz.
[b] Based on the fact that 200 sequences threaded through 1000 structures take 1 CPU day on an SGI Origin 200.
[c] Refinement takes 1 CPU day per sequence on an SGI Origin 200.

The resulting set of predicted structures could be used for molecular f assignment [12, 23]. In addition, since it is not practical to determ structure of all proteins in a genome, some choices have to be made. E evolutionary based folding, putative novel folds could be identified, an guide experimental studies by suggesting likely targets [84, 166, 16 Using either this approach or threading, such models could provid starting structures for NMR refinement. Furthermore, using the exact

methods described in Section 2, low-resolution models of single domain proteins could be produced from a limited amount of experimental data [72, 86, 164, 165]. These models could then be used to help refine the structures, thereby speeding up the process of structure determination. Thus, the field protein structure prediction is likely to play a vital role in the genomics revolution.

7. OUTLOOK

The last decade has seen considerable progress in the field of protein structure prediction using reduced protein models. At present, for small single domain proteins quite often it is possible to identify a handful of folds, one being native [166, 167, 172, 204]. On the other hand, improvements in the energy functions that select the native structure and better sampling techniques must be developed. Such approaches will allow for the identification of the native topology with greater certainty and the extension of these approaches to larger, single domain proteins. This is likely to be accomplished by the convergence of sequence based, threading and ab initio folding approaches. Clearly, evolutionary information can provide tremendous structural insights, the key question is how to extract this information in the appropriate manner. Similarly, local fragment threading can provide a set of effective building blocks from which to assemble novel native folds. While the latter has been previously tried [205], it has failed due to the lack of an adequate energy function; the requisite enhancement can be provided using evolutionary information. One is still left with the problem of conformational sampling, and here distance geometry may play a very important role [164]. Furthermore, in the near future, one is likely to see the development and maturation of techniques that can bring low resolution models closer to the native structure. Better atomic potentials, better sampling and/or the use of evolutionary information may achieve such model enhancement in detailed as opposed to reduced models.

Improvements in the ability to predict tertiary structure will allow these techniques to assist in the rapid determination of protein structures. Such tools are necessary if structural genomics, which is designed to determine the tertiary structure of all proteins [220-222], is to make major progress. By suggesting proteins of novel fold and/or function, tertiary structure prediction can be used to prioritize the selection of proteins whose structure will determined by experiment. It will also assist in the rapid determination of the structure of such proteins. Thus, the outlook for the future of the field of protein s prediction is very bright. Not only will considerable theoretical prog made in the near future, but also the practical applications of these tec can make them an important contributor to the genomics revolution.

REFERENCES

1. C. Bult, *et al.*, Science, 273 (1996) 1058-1073.
2. W.C. Barker, *et al.*, Nucleic Acids Res., 26 (1998) 27-32.
3. F.R. Blattner, *et al.*, Science, 277 (1997) 1453-1462.
4. R.D. Fleischmann, *et al.*, Science, 269 (1995) 496-512.
5. C.M. Fraser, *et al.*, Science, 270 (1995) 397-403.
6. R. Gibbs, Nature Genet., 11 (1995) 121-125.
7. T. Kaneko, *et al.*, DNA Res., 3 (1996) 109-136.
8. F. Kunst, *et al.*, Nature, 390 (1997) 249-256.
9. H.W. Mewes, *et al.*, Nature, 387 (1997) 7-65.
10. J.-F. Tomb, *et al.*, Nature, 388 (1997) 539-547.
11. E. Koonin, R. Tatusov, and M.Y. Galperin, Curr. Opin. Struct. I (1998) 355-363.
12. J.S. Fetrow and J. Skolnick, J. Mol. Biol., 281 (1998) 949-968.
13. S.S. Sturrock and J.F. Collins, *MPsrch version 1.3*, in *Biocor Research Unit*. 1993: University of Edinburgh, U.K.
14. W.R. Pearson and D.J. Lipman, Proc. Nat. Acad. Sci. USA, 85 (1988 2448.
15. S.F. Altschul, *et al.*, Nucleic Acids Res., 25 (1997) 3389-3402.
16. S. Henikoff and J.G. Henikoff, Nucleic Acids Res., 19 (1991) 6565-6
17. S. Henikoff and J.G. Henikoff, Genomics, 19 (1994) 97-107.
18. A. Bairoch, P. Bucher, and K. Hofmann, Nucleic Acids Res., 24 189-196.
19. R. Abagyan and S. Batalov, J. Mol. Biol., 273 (1997) 355-368.
20. P. Bork and E.V. Koonin, Nature Genet., 18 (1998) 313-318.
21. E.V. Koonin, Curr. Biol., 7 (1997) R656-R659.
22. J.S. Fetrow, A. Godzik, and J. Skolnick, Protein Sci., (1998) submitte
23. J.S. Fetrow, A. Godzik, and J. Skolnick, J. Mol. Biol., (1998) accep publication.
24. B.R. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan, Karplus, J. Comp. Chem., 4 (1983) 187-217.

25.D.A. Pearlman, *et al.*, *AMBER*, . 1991, University of California: San Francisco.
26.R. Friesner and M. Beachy, Curr. Opin. Struct. Biol., 8 (1998) 157-262.
27.H.A. Scheraga, M.-H. Hao, and J. Kostrowicki, *Theoretical studies of protein folding*, in *Methods in Protein Structure Analysis*, M.Z. Atassi and E. Appela, Editors. 1995, Plenum Press: New York.
28.D.A. Hinds and M. Levitt, Proc. Natl. Acad. Sci. USA, 89 (1992) 2536-2540.
29.M.J. Sippl and S. Weitckus, Proteins, 13 (1992) 258-271.
30.H. Flockner, M. Braxenthaler, P. Lackner, M. Jaritz, M. Ortner, and M.J. Sippl, Proteins, 23 (1996) 376-386.
31.S. DeBolt and J. Skolnick, Prot. Engn., 9 (1996) 637-655.
32.M. Karplus and G.A. Petsko, Nature, 347 (1990) 631-639.
33.A. Rey and J. Skolnick, Chemical Physics, 158 (1991) 199-219.
34.E.W. Knapp and A. Irgens-Defregger, J. Comput. Chem., 14 (1993) 19-29.
35.A. Kolinski and J. Skolnick, Proteins, 18 (1994) 338-352.
36.W.R. Krigbaum and S.F. Lin, Macromolecules, 15 (1982) 1135-1145.
37.D.G. Covell, Proteins, 14 (1992) 409-420.
38.E. Shakhnovich, G. Farztdinov, and A.M. Gutin, Phys. Rev. Lett., 67 (1991) 1665-1668.
39.M.-H. Hao and H.A. Scheraga, J. Chem. Phys, 102 (1995) 1334-1348.
40.D. Hoffmann and E.W. Knapp, Phys. Rev. E, 53 (1996) 4221-4224.
41.A.F.P. de Araujo and T.C. Pochapsky, Folding & Design, 1 (1996) 299-314.
42.M.-H. Hao and H.A. Scheraga, J. Phys. Chem, 98 (1994) 4940-4948.
43.M.-H. Hao and H.A. Scheraga, J. Phys. Chem., 98 (1994) 9882-9893.
44.U.H.E. Hansmann and Y. Okamoto, J. Comput. Chem., 14 (1993) 1333-1338.
45.A. Kolinski, W. Galazka, and J. Skolnick, Proteins, 26 (1996) 271-287.
46.J. Kostrowicki, L. Piela, B.J. Cherayil, and H.A. Scheraga, J. Phys. Chem., 95 (1991) 4114.
47.L. Piela, J. Kostrowicki, and H.A. Scheraga, J. Phys. Chem., 93 (1989) 3339-3346.
48.T. Dandekar and P. Argos, J. Mol. Biol., 236 (1994) 844-861.
49.T. Dandekar and P. Argos, J. Mol. Biol., 256 (1996) 645-660.
50.M. Levitt, Current Opinion in Structural Biology, 1 (1991) 224-229.
51.V. Daggett and M. Levitt, J. Mol. Biol., 232 (1993) 600-619.
52.V. Daggett and M. Levitt, Curr. Opin. Struct. Biol., 4 (1994) 291-295.
53.C.L.I. Brooks, Curr. Opin. Struct. Biol., 3 (1993) 92-98.
54.C.L.I. Brooks, M. Karplus, and B.M. Pettitt, *Proteins: A the perspective of dynamics structure and thermodynamics*. Adva Chemical Physics. 1988, New York: Wiley.
55.J. Hirst, D. and C. Brooks III, L., J. Mol. Biol., 34 (1995) 7614-21.
56.P.A. Kollman, Acc. Chem. Res., 29 (1996) 461-469.
57.Y. Duan, L. Wan, and P. Kollman, Proc. Natl. Sci.,USA, 95 (1998 9902.
58.A. Godzik, A. Kolinski, and J. Skolnick, J. Comp. Chem., 14 (1993 1202.
59.A. Kolinski and J. Skolnick, Proteins, 18 (1994) 353-366.
60.A. Kolinski, W. Galazka, and J. Skolnick, J. Chem. Phys., 103 10286-10297.
61.A. Kolinski, M. Milik, J. Rycombel, and J. Skolnick, J. Chem. Ph (1995) 4312-4323.
62.B.H. Park and M. Levitt, J. Mol. Biol., 249 (1995) 493-507.
63.J. Skolnick, A. Kolinski, C. Brooks III, A. Godzik, and A. Rey, Cur 3 (1993) 414-423.
64.D.G. Covell and R.L. Jernigan, Biochemistry, 29 (1990) 3287-3294.
65.N. Go and H. Taketomi, Proc. Natl. Acad. Sci. USA, 75 (1978) 559-
66.N. Go, H. Abe, H. Mizuno, and H. Taketomi, eds. *Protein folding*. Jaenicke. 1980, Elsevier/North Holland: Amsterdam. 167-181.
67.J. Skolnick and A. Kolinski, Science, 250 (1990) 1121-1125.
68.M.-H. Hao, S. Rackovsky, A. Liwo, M.R. Pinkus, and H.A. Scherag Natl. Acad. Sci. USA, 89 (1992) 6614-6618.
69.M. Levitt, J. Mol. Biol., 104 (1975) 59-107.
70.G.M. Crippen, Biochemistry, 30 (1991) 4232-4237.
71.M. Vieth, A. Kolinski, I. Brooks, C. L., and J. Skolnick, J. Mol. Bi (1995) 448-467.
72.J. Skolnick, A. Kolinski, and A.R. Ortiz, J. Mol. Biol., 265 (1997) 21
73.A. Liwo, *et al.*, J. Protein Chem., 13 (1994) 375-380.
74.A. Liwo, M.R. Pincus, R.J. Wawak, S. Rackovsky, and H.A. S Protein Sci., 2 (1993) 2725-1731.
75.M. Levitt and A. Warshel, Nature, 253 (February 27)(1975) 694-698
76.Y. Ueda, H. Taketomi, and N. Go, Biopolymers, 17 (1978) 1531-15
77.K.A. Dill, *et al.*, Prot. Sci., 4 (1995) 561-602.

78.E.I. Shakhnovich and A.M. Gutin, Proc. Natl. Acad. Sci. USA, 90 (1993) 7195-7199.
79.A. Sali, E. Shakhnovich, and M. Karplus, J. Mol. Biol., 235 (1994) 1614-1636.
80.H.S. Chan and K.A. Dill, Macromolecules, 22 (1989) 4559-4573.
81.V.I. Abkevich, A.M. Gutin, and E.I. Shakhnovich, J. Chem. Phys., 101 (1994) 6052-6062.
82.P.G. Wolynes, *The basics of protein folding physics*, in *Physics of Biomaterials: Fluctuations, Self-assembly and Evolution*, T. Riste and D. Sherrington, Editors. 1996, Kluwer Academic Publ.: Netherlands. p. 235-248.
83.M. Vieth, A. Kolinski, C.L. Brooks III, and J. Skolnick, J. Mol. Biol., 237 (1994) 361-367.
84.C. Simmerling, M. Lee, A.R. Ortiz, A. Kolinski, J. Skolnick, and P.A. Kollman, J.A.C.S., submitted (1998) .
85.A. Kolinski and J. Skolnick, *Lattice models of protein folding, dynamics and thermodynamics*. 1996, Austin, TX.: R. G. Landes. 200.
86.A. Kolinski and J. Skolnick, Proteins, 32 (1998) 475-94.
87.R.C. Brower, G. Vasmatiz, M. Silverman, and C. Delisi, Biopolymers, 33 (1993) 329-334.
88.E.E. Lattman, K.M. Fiebig, and K.A. Dill, Biochemistry, 33 (1994) 6158-6166.
89.K.F. Lau and K.A. Dill, Macromolecules, 22 (1989) 3986-3997.
90.P.D. Thomas and K.A. Dill, J. Mol. Biol., 257 (1996) 457-469.
91.B. Honig and F.E. Cohen, Folding & Design, 1 (1996) R17-R20.
92.M. Hao and H.A. Scheraga, J. Mol. Biol., (1998) in press.
93.V.I. Abkevich, A.M. Gutin, and E.I. Shakhnovich, Protein Sci., 4 (1995) 1167-1177.
94.V.I. Abkevich, A.M. Gutin, and E.I. Shakhnovich, Folding & Design, 1 (1996) 221-230.
95.R.S. DeWitte and E.I. Shakhnovich, Protein Sci., 3 (1994) 1570-1581.
96.A.V. Finkelstein and E.I. Shakhnovich, Biopolymers, 29 (1989) 1681-1694.
97.M. Karplus and E. Shakhnovich, *Protein folding: Theoretical studies of thermodynamics and dynamics*, in *Protein Folding*, T.E. Creighton, Editor. 1992, W.H.Freeman. p. 127-196.
98.L.A. Mirny, V. Abkevich, and E.I. Shakhnovich, Folding & Design, 1 (1996) 103-116.
99.A. Sali, E. Shakhnovich, and M. Karplus, Nature, 369 (1994) 248-25
100.E.I. Shakhnovich and A.V. Finkelstein, Biopolymers, 28 (1989 1680.
101.E.I. Shakhnovich and A.M. Gutin, Biophys. Chem., 34 (1989) 187-
102.E. Shakhnovich, I. and A. Finkelstein, V., Biopolymers, 26 (1989 1694.
103.E.I. Shakhnovich and A.M. Gutin, Protein Engng., 6 (1993) 793-80
104.E.I. Shakhnovich, Phys. Rev. Lett., 72 (1994) 3907-3910.
105.E.I. Shakhnovich, Folding & Design, 1 (1996) R50-R54.
106.E. Shakhnovich, Curr. Opin. Struct. Biol., 7 (1997) 29-40.
107.H.S. Chan and K.A. Dill, J. Chem. Phys., 95 (1991) 3775-3787.
108.C.J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. USA, 9(6369-6372.
109.E. O'Toole, R. Venkataramani, and A.Z. Panagiotopoulous, AICh (1995) 954-958.
110.C.J. Camacho and D. Thirumalai, Proteins, 22 (1995) 27-40.
111.J.D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. U (May)(1990) 3526-3529.
112.J.D. Honeycutt and D. Thirumalai, Biopolymers, 32 (1992) 695-70
113.P.G. Wolynes, *The basics of protein folding physics*, in *Ph Biomaterials: Fluctuations, Self assembly and evolution*, T. Riste Sherrington, Editors. 1996, Kluwer Academic Publ.: Netherlands. 248.
114.K. Dill and H. Chan, Nature Struct. Biol., 4 (1997) 1-19.
115.A. Sikorski and J. Skolnick, Biopolymers, 28 (1989) 1097-1113.
116.A. Sikorski and J. Skolnick, Proc. Natl. Acad. Sci. USA, 86 (198 2672.
117.A. Sikorski and J. Skolnick, J. Mol. Biol., 212 (1990) 819-836.
118.A. Kolinski, M. Milik, and J. Skolnick, J. Chem. Phys., 94 (199 3985.
119.A. Kolinski and J. Skolnick, J. Phys. Chem., 97 (1992) 9412-9426.
120.J. Skolnick and A. Kolinski, Annu. Rev. Phys. Chem., 40 (1989) 2
121.J. Skolnick, A. Kolinski, and R. Yaris, Biopolymers, 28 (1989) 10
122.J. Skolnick, A. Kolinski, and R. Yaris, Proc. Natl. Acad. Sci. (1989) 1229-1233.
123.J. Skolnick and A. Kolinski, J. Mol. Biol., 212 (1989) 787-817.

124.J. Skolnick, A. Kolinski, and A. Sikorski, Chemical Design Automation News, 5 (1990) 1-20.
125.J. Skolnick and A. Kolinski, J. Mol. Biol., 221 (1991) 499-531.
126.H. Taketomi, F. Kano, and N. Go, Biopolymers, 27 (1988) 527-559.
127.H. Taketomi, Y. Ueda, and N. Go, Int. J. Pept. Protein Res., 7 (1988) 445-449.
128.J.U. Bowie, R. Lüthy, and D. Eisenberg, Science, 253 (1991) 164-170.
129.S.H. Bryant and C.E. Lawrence, Proteins, 16 (1993) 92-112.
130.A.V. Finkelstein and B.A. Reva, Nature, 351 (1991) 497-499.
131.A. Godzik and J. Skolnick, Proc. Natl. Acad. Sci. USA, 89 (1992) 12098-12102.
132.L. Jaroszewski, L. Rychlewski, B. Zhang, and A. Godzik, Protein Sci., 7 (1998) 1431-1440.
133.D.T. Jones, W.R. Taylor, and J.M. Thornton, Nature, 358 (1992) 86-89.
134.R. Lathrop and T.F. Smith, J. Mol. Biol., 255 (1996) 641-665.
135.T. Madej, J.F. Gibrat, and S.H. Bryant, Proteins, 23 (1995) 356-369.
136.V.N. Maiorov and G.M. Crippen, J. Mol. Biol., 277 (1992) 876-888.
137.C. Ouzounis, C. Sander, M. Scharf, and R. Schneider, J. Mol. Biol., 232 (1993) 805-825.
138.R.B. Russel, R.R. Copley, and G.J. Barton, J. Mol. Biol., 259 (1996) 349-365.
139.R. Thiele, R. Zimmer, and T. Lengauer, ISMB, 3 (1995) 384-392.
140.M. Wilmanns and D. Eisenberg, Protein Eng., 8 (1995) 626-639.
141.A. Sali, J.P. Overington, M.S. Johnson, and T.L. Blundell, TIBS, 15 (1990) 235-250.
142.A. Sali and T. Blundell, J. Mol. Biol., 234 (1993) 779-815.
143.C. Chothia and A. Finkelstein, Annu. Rev. Biochem., 59 (1990) 1007-39.
144.A. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia, J. Mol. Biol., 247 (1995) 536-540.
145.UCLA, *The UCLA-DOE benchmark to assess the performance of fold recognition methods*, . 1996: Los Angeles.
146.N.N. Aleksandrov, R. Nussinov, and R.M. Zimmer. *Fast fold recognition via sequence to structure alignment and contact capacity potentials*. in *Pacific Symposium on Biocomputing, 96*. 1996. Hawaii: World Scientific.
147.L. Jaroszewski and A. Godzik, Protein Eng., (1996) submitted.
148.A. Godzik, A. Kolinski, and J. Skolnick, J. Comp. Aided Mol. Design, 7 (1993) 397-438.
149.R.T. Miller, D.T. Jones, and J.M. Thornton, FASEB, 10 (1996) 17
150.T. Smith, L. Lo Conte, J. Bienkowska, C. Gaitatzes, R. Rogers Jr Lathrop, J. Comp. Biol., 4 (1997) 217-25.
151.A. MarchlerBauer and S.H. Bryant, Trends Biochem. Sci., 22 (19 240.
152.A. MarchlerBauer, M. Levitt, and S. Bryant, Proteins. Suppl., 1 (1 82.
153.C. Lemer, M. Rooman, and S. Wodak, Proteins, 23 (1996) 337-35.
154.F.E. Cohen, A.R. M., I.D. Kuntz, and R.J. Fletterick, Biochem (No. 21)(1983) 4894-4904.
155.I.P. Crawford, T. Niermann, and K. Kirschner, Proteins, 2 (1987)
156.O.B. Ptitsyn and A.A. Rashin, Biophys. Chem., 3 (1975) 1-20.
157.A. Warshel and M. Lewitt, J. Mol. Biol., 106 (1976) 421-437.
158.M.J. Smith Brown, D. Kominos, and R.M. Levy, Prot. Engn., 605-614.
159.G. Chelvanayagam, L. Knecht, T. Jenny, S.A. Benner, and G.H. Foding & Design, 3 (1998) 149-160.
160.C. Wilson and S. Doniach, Proteins, 6 (1989) 193-209.
161.K.A. Dill and K. Yue, Prot. Sci., 5 (1996) 254-261.
162.D. Eisenberg and A.D. McLauchlan, Nature, 319 (1986) 199-203.
163.S. Miyazawa and R.L. Jernigan, Macromolecules, 18 (1985) 534-5
164.C. Mumenthaler and W. Braun, Prot. Sci., 4 (1995) 863-871.
165.A. Aszodi, M.J. Gradwell, and W.R. Taylor, J. Mol. Biol., 251 (19 326.
166.A.R. Ortiz, A. Kolinski, and J. Skolnick, J. Mol. Biol., 277 (19 448.
167.A.R. Ortiz, A. Kolinski, and J. Skolnick, Proteins Struct. Funct. (1998) 287-294.
168.B. Rost and C. Sander, J. Mol. Biol., 232 (1993) 584-599.
169.U. Gobel, C. Sander, R. Schneider, and A. Valencia, Proteins, 1 309-317.
170.O. Olmea and A. Valencia, Folding & Design, 2 (1997) S25-S32.
171.J. Skolnick, Proteins, submitted (1998) .
172.A.R. Ortiz, A. Kolinski, and J. Skolnick, Proc. Natl. Sci. USA, 9 1020-1025.

173.J. Skolnick and A. Kolinski, *Protein modelling*, in *Encyclopedia of Computational Chemistry*, P. von Rauge Schleyer, Editor. 1997, John Wiley & Sons: New York.
174.M. Levitt, J. Mol. Biol., 104 (1976) 59-107.
175.A. Wallqvist and M. Ullner, Proteins, 18 (1994) 267-289.
176.V.N. Maiorov and G.M. Crippen, Proteins, 20 (1994) 167-173.
177.A. Monge, E.J.P. Lathrop, J.R. Gunn, P.S. Shenkin, and R.A. Friesner, J. Mol. Biol., 247 (1995) 995-1012.
178.S. Sun, Protein Sci., 2 (1993) 762-785.
179.A. Kolinski, A. Godzik, and J. Skolnick, J. Chem. Phys., 98 (1993) 7420-7433.
180.A. Kolinski, W. Galazka, and J. Skolnick, J. Chem. Phys., 108 (1998) 2608-2617.
181.A. Godzik, A. Kolinski, and J. Skolnick, Protein Sci., 4 (1995) 2107-2117.
182.A. Godzik, Current Biol., 4 (1996) 363-366.
183.J. Skolnick and A. Kolinski, *Monte Carlo lattice dynamics and the prediction of protein folds*, in *Computer simulations of biomolecular systems. Theoretical and experimental studies.*, W.F. van Gunsteren, P.K. Weiner, and A.J. Wilkinson, Editors. 1996, ESCOM Science Publ.
184.J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik, Prot. Sci., 6 (1997) 676-688.
185.R.L. Jernigan and I. Bahar, Curr. Opin. Struct. Biol., 6 (1996) 195-209.
186.K. Wutrich, *NMR of proteins and nucleic acids*. 1986, New York: Wiley.
187.L. Regan, A. Rockwell, Z. Wasserman, and W. De Grado, Prot. Sci., 3 (1994) 2419-27.
188.A. Monge, R.A. Friesner, and B. Honig, Proc. Natl. Acad. Sci. USA, 91 (1994) 5027-5029.
189.G.J.R. Gunn, A. Monge, and R.A. Friesner, J. Phys. Chem., 98 (1994) 702-711.
190.R.A. Friesner and J.R. Gunn, Annu. Rev. Biophys. Biomol. Struct., 25 (1996) 315-342.
191.T.F. Havel and K. Wuthrich, J. Mol. Biol., 182 (1985) 281-294.
192.A. Godzik, J. Skolnick, and A. Kolinski, Proc. Natl. Acad. Sci. USA, 89 (1992) 2629-2633.
193.M.J. Bayley, G. Jones, P. Willett, and M.P. Williamson, Protein Sci., 7 (1998) 491-499.
194.A. Aszodi and W.R. Tylor, Foding & Design, 1 (1996) 325-34.
195.B. Cheng, A. Nayeem, and H.A. Scheraga, J. Comput. Chem., 1' 1453-1480.
196.K.A. Olszewski, A. Kolinski, and J. Skolnick, Proteins, 25 (19' 299.
197.R. Samudrala, Y. Xia, M. Levitt, and E. Huang, *A combinded appr ab initio construction of low resolution tertiary structures*, in *Symposium on biocomputing'99*, K.D. RB Altman, L. Hunter, T. I Editor. 1998, World Scientific: Singapore. p. in press.
198.D. Qui, P. Shenkin, F. Hollinger, and W. Still, J. Phys. Chem., 10 3005-3014.
199.X. Duara, B. Jaun, D. Seebach, W.v. Gunsteren, and A. Mark, Biol., 280 (1998) 925-932.
200.A.T. Hagler and B. Honig, Proc. Natl. Acad. Sci. USA, 75 (1978) :
201.J.T. Pedersen and J. Moult, J. Mol. Biol., 269 (1997) 240-259.
202.F. Avgelj and L. Fele, Proteins, 31 (1998) 74-96.
203.Y. Cui, R. chen, and W. Wong, Proteins, 31 (1998) 247-257.
204.D. Jones, Proteins,Suppl, 1 (1997) 185-191.
205.K.T. Simons, C. Klooperberg, E. Huang, and D. Baker, J. Mol. B (1997) 209-225.
206.B. Rost and C. Sander, Proteins, 19 (1994) 55-72.
207.B. Rost and C. Sander, Proteins, 23 (1996) 295-300.
208.D. Jones, Curr Opin Struct. Biol., 7 (1997) 377-387.
209.B. Rost and C. Sander, Ann. Rve. Biophys. Biomol. Struct., 2' 113-136.
210.A.R. Ortiz, W.-P. Hu, A. Kolinski, and J. Skolnick, Science submitted.
211.B. Rost and C. Sander, J. Mol. Biol., 232 (1993) 584-599.
212.A. Kolinski, J. Skolnick, A. Godzik, and W.P. Hu, Proteins, 2' 290-308.
213.L. Holm and C. Sander, Nucleic Acids Res, 25 (1997) 231-234.
214.A. Ortiz and J. Skolnick, J. Mol. Biol., submitted (1998) .
215.L. Mirny, V.I. Abkevich, and E. Shakhnovich, Proc. Natl. Acad. S 95 (1998) 4976-4981.
216.D.T. Jones, Curr. Opin. Struct. Biol., 7 (1997) 377-387.
217.I. Radhakrishnan, G.C. Perez_Alvarado, D. Parker, H.J. Dyso Montminy, and P.E. Wright, Cell, 91 (1997) 741-752.

218.U. Essman, L. Perera, M. Berkowitz, T. Darden, H. Lee, and L. Pedersen, J. Chem. Phys, 103 (1995) 8577-8593.
219.A. Kolinski, R. Rotkiewicz, B. Ilkowski, and J. Skolnick, J. Mol. Biol., submitted (1998) .
220.L. Shapiro and C. Lima, Structure, 6 (1998) 265-267.
221.S. Kim, Nat. Struct. Biol., 5 (1998) 643-645.
222.T. Gaasterland, Trends in Genetics, 14 (1998) 135.