PROTEIN STRUCTURE PREDICTION II

Jeffrey Skolnick[1 ,2]
Yang Zhang[1]

Because the molecular function of a protein depends on its three dimensional structure, which is often unknown, protein structure prediction is an essential tool in proteomics. The state of the art of protein structure prediction is reviewed with emphasis on knowledge-based comparative modeling/threading approaches that exploit the observation that the protein data bank (PDB) is complete for low-resolution, single domain proteins. The recently developed *TASSER* structure prediction algorithm is described; its ability to produce structures closer to the native state than to the initial template structure demonstrated, along with encouraging results for membrane protein tertiary structure prediction and its ability to predict NMR quality structures in 1/5 of the cases. The quality of predicted structures required for the inference of biochemical function and describe structure-based approaches to predict protein-protein interactions is also examined. Applications to a number of proteomes are presented. The weaknesses and strengths of existing approaches are summarized with an emphasis on the importance of large scale benchmarking. Finally, the outlook for future progress is reviewed.

INTRODUCTION

Over the past decade, the success of genome sequence efforts has brought about a paradigm shift in biology [1]. There is increasing emphasis on the large-scale, high-throughput examination of all genes and gene products of an organism, with the aim of assigning their functions[2]. Of course, biological function is multifaceted, ranging from molecular/biochemical to cellular or physiological to phenotypical [3]. In practice, knowledge of the DNA sequence of an organism and the identification of its open reading frames (ORFs) does not directly provide functional insight. Here, the focus is on the proteins in a genome, *viz*. the proteome, but recognize that proteins are only a subset of all biologically important molecules and address aspects of molecular/biochemical function and protein-protein interactions. At present, evolutionary based approaches can provide insights into some features of the biological function of about 40-60% of the ORFs in a given proteome [4]. However, pure evolutionary based approaches increasingly fail as the protein families become more distant [5], and predicting the functions of the unassigned ORFs in a genome remains an important challenge. Because the biochemical function of a protein is ultimately determined by both the identity of the functionally important residues and the three dimensional structure of the functional site, protein structures represent an essential tool in annotating genomes [6-11]. The recognition of the role that structure can play in elucidating function is one impetus for structural genomics that aims for high-throughput

[1] Center of Excellence in Bioinformatics, University at Buffalo, Buffalo, NY 14203.
[2] skolnick@buffalo.edu, Phone (716) 849-6711

protein structure determination [12]. Another is to provide a complete library of solved protein structures so that an arbitrary sequence is within modeling distance of an already known structure[13]. Then, the protein folding problem, *viz.* the prediction of a protein's structure from its amino acid sequence, could be solved by enumeration. In practice, the ability to generate accurate models from distantly related templates will dictate the number of protein folds that need to be determined experimentally [14-16]. Protein-protein interactions, which are involved in virtually all cellular processes [17], represent another arena where protein structure prediction could play an important role. This area is in ferment, with considerable concern about the accuracy and consistency of high throughput experimental methods [18].

In what follows, an overview of areas that comprise the focus of this chapter is presented. First, the state-of-the-art of protein structure prediction is discussed. Then, the status of approaches to biochemical function prediction based on both protein sequence and structure is reviewed, and followed by a review of the status of approaches for determining protein-protein interactions. Then, some recent promising advances in these areas are described. In the conclusion section, the status of the field and the directions for future research are summarized.

BACKGROUND

Historically, protein structure prediction approaches are divided into the three general categories, Comparative Modeling (*CM*)[19], threading[20] and New Fold methods or *ab initio* folding[21-23], that are schematically depicted in Figure 1. In *CM*, the protein's structure is predicted by aligning the target protein's sequence to an evolutionarily related template sequence with a solved structure in the PDB [24]; i.e. two homologous sequences are aligned, and a three dimensional model built based on this alignment [25]. In threading, the goal is to match the target sequence whose structure is unknown to a template that adopts a known structure, whether or not the target and template are evolutionarily related [26]. It should identify analogous folds *viz.* where they adopt a similar fold without an apparent evolutionary relationship [27-29]. Note that the distinction between these approaches is becoming increasingly blurred [29-31]. Certainly, the general approach of *CM* and threading is the same: Identify a structurally related template, identify an alignment between the target sequence and the template structure, build a continuous, full length model, and then refine the resulting structure [26]. *Ab initio* folding usually refers to approaches that model protein structures on the basis of physicochemical principles. However, many recently developed New Fold/*ab initio* approaches often exploit evolutionary and threading information [30] (e.g. predicted secondary structure or contacts), although some versions are more physics-based [32], perhaps such approaches such be referred to as semi-first principles. Indeed, a number of groups have developed approaches that span the range from *CM* to *ab initio* [29, 30] folding that performed reasonably well in CASP5, the fifth biannual community wide experiment to assess the status of the field of protein structure prediction[33].

COMPARATIVE MODELING

Comparative modeling can be used to predict the structure of those proteins whose sequence identity is above 30% with a template protein sequence [34], although progress has been reported at lower sequence identity [26]. An obvious limitation is that it requires a homologous protein, the template, whose structure is known. When proteins have more than 50% sequence identity to their templates, in models built by *CM* techniques, the backbone atoms [19] can have up to a 1 Å root-mean-square-deviation (RMSD) from native; this is comparable to experimental accuracy [9]. For target proteins with 30% to 50% sequence identity to their templates, the backbone atoms often have about 85% of their core regions within a RMSD of 3.5 Å from native, with errors mainly in the loops [19]. When the sequence identity drops below 30%, the model accuracy by *CM* sharply decreases because of the lack of significant template hits and substantial alignment errors. The sequence identity <30% is usually termed the "twilight" zone for sequence-based alignment and more than half of genome sequences are at these distances to known proteins in PDB. For all sequence identity ranges, the predicted structures are generally closer to the template on which they are based rather than to their native conformation[34]. This was true in the recent CASP5 protein structure prediction experiment [35]. Another issue is the accurate construction of the loops. While progress has been made for short loops [36], for longer loops, significant problems remain [35]. Therefore, it is essential to develop an automated technology that can deal with proteins in the twilight zone of sequence identity, then build models that are closer to the native structure than to the template on which they are based [37, 38]. Many recent developed threading algorithms start to be able to identify structural analogs in the twilight zone. But little progress has been reported with regard to the template refinements.

Despite these limitations, *CM* has been applied to predict the tertiary structure of the ORFS in a number of proteomes [39]. At present, about 40-50% of all sequences have a homologous protein of known structure, with *CM* results compiled in the *PEDANT* [40], *GTOP* [41], *MODBASE* [42], and *FAMS* [41] databases. This percentage is slowly increasing as new structures are being solved at an increasing rate. Interestingly, most newly solved structures exhibit an already known fold [16], an issue examined below.

THREADING

The formulation of a threading algorithm involves three choices: First, the interaction sites must be chosen. Due to computational complexity, these are taken to be a subset of the protein's heavy atoms and can be the Cαs [43], Cβs [44], side-chain centers of mass [45], specially defined interaction centers [46], or any side-chain atom[47]. Second, the functional form of the energy is chosen, with examples ranging from contact [47] to continuous distance-dependent potentials [44]. The energy can include predicted secondary structure preferences [48] or burial patterns [49]. To improve both template recognition ability and the quality of the alignment, most successful threading approaches combine both sequence and structural information [27, 48, 50]. Third, given an energy function, the optimal alignment of the target sequence to each structural template must be found. If the "energy" terms are local (e.g. secondary structure propensities and/or sequence profiles), then dynamic programming [51] is best. If pair interactions are considered (which use a non local scoring function), the interactions in the template structure must be updated to reflect the target sequence. Some approaches employ dynamic programming with a frozen environment (with interaction partners taken from the template protein)[20], followed by iterative updating[47]; others employ double dynamic programming that updates some interactions recognized as being the most important in the first pass of dynamic programming [52]. Other computationally more intensive variants include the actual partners in the target sequence and use Monte Carlo [53] or branch-and-bound search strategies [54]. A reasonably successful and faster alternative uses a sequence profile to align the target sequence to the template structure; then, the partners in the target sequence are used to evaluate the pair interactions [45, 50]. These approaches suffer from the disadvantages that the template structure never adjusts to reflect modifications due to differences in the target and template sequences, and one cannot do better than the best structural alignment between the template and target structures [16 , 55, 56].

As demonstrated in CASP5 [29, 30, 57-60], there are now a number of threading methods that significantly outperform sequence-only approaches such as *PSI-BLAST*[58]. Examples include *PROSPECT II* [27], *GENTHREADER* [48] and *PROSPECTOR* [50]. These algorithms found some analogous[29] structural templates for targets in the fold recognition/analogous (FR/A) category [5]. However, threading had many outstanding issues in common with *CM:* the need to improve aligned regions and move them closer to the native structure than the initial template alignment and the need to have a good loop building algorithm that fills in the gapped region and generates statistically significant loop predictions. Furthermore, selection of the best model was often problematic [29].

METAPREDICTOR BASED APPROACHES

CASP5/CAFASP3 demonstrated the power of *Metapredictors* (defined as automated predictors that combine consensus information from a variety of threading and sequence based servers to make more accurate consensus structural predictions) such as *3-D SHOTGUN* [59], *PCONS* [60] and *ROBETTA* [61] that gave results competitive with the best human predictors [59]. *3D-SHOTGUN* and *PCONS* [60] do not simply select a model from the input models, but generate more complete and accurate hybrid models by splicing fragments from the individual models; however, these can have steric clashes, sometimes making the construction of physically realistic models impossible. Nonetheless, based on *EVA* [62] and *LiveBench* [63] results, *Metapredictors* are quite promising. For example, in large scale testing, *3D-SHOTGUN* produced models up to 28% higher Maxsub score than any of the individual methods and had 17% higher specificity than any individual method. Here, the specificity is defined as the number of correct predictions with confidence score higher than the first false prediction. These results illustrate the potential power of the *Metaprediction* approach. However, the ultimate success of *Metaprediction* lies in the underlying accuracy of the individual contributing servers.

COMPLETENESS OF THE PDB

CM/threading approaches cannot succeed if a structure related to the target sequence is not already solved. Therefore, the key issue for their applicability is the completeness of the *PDB* [24]. One way to explore this issue is to use structural alignment algorithms (which find the best structural match between a pair of proteins where the labeling of residues to be matched is not fixed in advance) to establish the structural relationship between newly solved protein structures and those already in the *PDB*. Indeed, the best alignment between a pair of protein structures that CM/threading can exploit is obtained from a structural alignment. One class of structural alignment algorithms employs dynamic programming [55] whose advantage is speed, but global optimality is not guaranteed. *DALI* [64] compares the intra-structural residue-residue distances in a pair of structures. Others [65, 66] compare spatial arrangements of secondary structure elements. Nussinov et al. [67] employ geometric hashing, while an incremental combinatorial extension (*CE*) method that combines structurally similar fragments was employed by Shindyalov & Bourne [68]. Kedem [69] defines the unit-vector RMS to detect chain segment similarities, and *MAMMOTH*[70] employs a heuristic algorithm to align low resolution structures and assigns their significance via an extreme value distribution.

Several authors compared a set of representative structures in the *PDB* [71] and emphasized the discreteness of structural space on the domain level of protein structures. On the other hand, using their *CE* method, Shindyalov & Bourne [72] recently pointed out that substructures obtained from an all-against-all structure comparison sometimes distribute among protein domains transgressing their respective fold types. These substructures are around 130 residue long, continuous chains, much longer than the conventional concept of supersecondary structure [73]. Harrison et al. also concluded that fold space is a continuum for some topology types in the β or α/β secondary structure class[74]. These studies suggest that there are rather large structure motifs of significant length that occur in many other folds. Yang & Honig[75] also showed that their structure comparison program detects structural similarity between different folds in the *SCOP* database[76]. This indicates that some regions of protein fold space are not as distinct as once thought.

Recently, using a more sensitive structure alignment algorithm, *SAL,* Kihara and Skolnick demonstrated that for low-to-moderate resolution structures, the *PDB* is essentially complete for single domain proteins [16]. That is, the global fold of essentially all single domain proteins can be found among the already solved structures in the *PDB*. Furthermore, protein structure space is very dense. The problem is to develop a threading algorithm that can find these related template structures/good alignments and build a model useful for functional annotation [77]. As shown below, there has been significant progress in this direction, but additional work needs to be done before the protein structure prediction problem can be viewed as being solved, at least by enumeration.

INFERENCE OF BIOCHEMICAL FUNCTION FROM STRUCTURE

Currently, most methods that assign the molecular/biochemical function of proteins are based on finding protein sequence homology [78] or conserved protein sequence or structural motifs [79-82] between the uncharacterized protein and a protein of known biochemical function. However, such methods often fail as the sequence identity drops below 40%. Because the global fold of a protein family is more conserved than its sequence, protein biochemical function prediction should benefit by the inclusion of structural information [38]. However, divergent and convergent evolution gives a non-unique relationship between function and fold. In general, fold type by itself is not sufficient for correct function prediction [83, 84], and additional information is required to infer biochemical function from structure. Several methods are based on three-dimensional descriptors of biologically relevant sites [7, 85-90]. In addition to active site descriptors characterizing the geometric features of catalytic residues [87], a number of approaches that describe binding sites focus on the conservation of geometrical arrangements of residues [90-93], the physicochemical properties of functional residues [90, 94], and/or ligand-cavity shape complementarity [95]. Many methods were specifically designed to recognize a particular type of ligand, e.g., adenylate [88], calcium [92] or DNA [94]; with more general methods only tested for a few ligand types [90, 91]. Of interest is the recently available *PINTS* (Patterns in Non-homologous Tertiary Structures) [31] approach designed to perform database searches against a collection of ligand-binding sites taken directly from *PDB* files [96].

Methods based on structural templates have been reasonably successful when applied to high-resolution structures. The question is what happens when predicted models of lower resolution are used? Given recent improvements in the performance of protein structure prediction algorithms [29, 77, 97], a structure-based method for protein function prediction that does not require high-resolution structures could be of practical value. The essential issue is to establish the quality of a predicted structure required to transfer a given biochemical function at a specified level of accuracy. In practice, the ability to detect functional sites in low-to-moderate resolution predicted structures had until recently only been tested for a few specific active site descriptors [92, 98]. Recently, Arakaki et al. have developed a method that automatically generates a structural library of 3D descriptors of enzyme active sites [77] (automated functional templates or *AFT*s; 593 in total for 162 different enzymes) based on functional and structural information extracted from public databases. The applicability to predicted structures was investigated by analyzing varying quality decoys derived from enzyme native structures. For 35% of decoys having a 3 to 4 Å backbone RMSD from the native structure, the *AFT* based method correctly identifies the active site and transfers the first three *EC* indices. A key challenge is to routinely generate predicted structures of at least this quality so that they can be used for biochemical function inference.

Approaches for determining protein-protein interactions

Given their biological importance[17], the development of efficient methods to detect and characterize protein-protein interactions and assemblies is a major theme of functional genomics and proteomics efforts [99]. Currently, two main types of experimental methods are used: (1). Yeast two-hybrid screening(*Y2H*) [100], which is mainly limited to binary interaction detection; and (2). The combination of large-scale affinity purification with Mass Spectrometry to detect and characterize multi-protein complexes [101]. First applied to yeast [102], these methods revealed the dense network of interactions linking proteins in the cell, but their error rate is high [18]. The coverage of *Y2H* screens seems incomplete, with many false negatives and false positives as evidenced by the limited overlap between sets of interacting proteins identified by different groups [103] and between those identified by *Y2H* and other approaches [104]. This discrepancy among experimental methods prompted keen interest in the development of computational methods for inferring protein-protein interactions [105-107]. Many consider protein-protein interactions in the most general context and often refer to "functionally interacting proteins" [106], implying that the proteins cooperate to carry out a given task without actually (or necessarily) engaging in physical contact. These methods exploit the fact that the genes of such cooperating proteins tend to be associated within genomes [107]. The earliest methods considered gene fusion [107, 108], conservation of gene order [109], and co-occurrence of genes in different genomes [107] as a means of inferring functional interactions. Subsequent methods frequently use protein sequence information and are based on the idea of gene co-evolution, which assumes that the genes of proteins that interact tend to evolve together [110]. Approaches based on the co-evolution model include phylogenetic tree topology comparison [111], gene preservation correlation, and correlated mutation approaches [110]. These methods offer several advantages - the idea of correlated evolution is *a priori* and fits basic biological principles. But their downside is their low signal-to-noise ratio [112]. Furthermore, methods based on a co-evolution model rely on the knowledge of the phylogenetic trees of the corresponding sets of proteins [113]. Given that the exact evolutionary path of a specific protein is unknown, one must infer phylogeny via a careful analysis of related proteins from different organisms, the so-called orthologs, whose identification is not straightforward [113]. Phylogenetic tree reconstruction is NP-complete [114]. Existing measures for assessing co-evolution (such as the Pearson correlation coefficient) attempt to avoid this problem by considering all protein homolog pairs. They are effective when the signal is strong, and often it is not.

A conceptually different set of methods uses information from protein quaternary structure, and deals more directly with the actual physical interactions between proteins. It is in this sense that protein-protein interactions are considered in what follows. These approaches not only suggest which two proteins interact, but also provide a quaternary structure. Salient examples of this class of approaches include promising extensions of homology modeling and threading techniques [115, 116] and neural net based approaches [117]. Below one promising extension of threading to treat predict protein-protein interactions is described [118, 119].

RECENT ADVANCES

CAN THE PROTEIN STRUCTURE PREDICTION PROBLEM BE SOLVED IN PRINCIPLE USING THE CURRENT PDB LIBRARY?

In recent studies, Skolnick et al constructed a representative set of all single domain proteins that have structures in the PDB (no two of which have more than 35% pairwise sequence identity to each other) ranging from 41 to 200 residues in length; there are 1489 such proteins [122] the *PDB200* benchmark set. Using an improved structural alignment algorithm, *SAL*, they then compared these proteins to a benchmark library that is no more than 20% identical to the target protein [77, 121]. The resulting average coverage and RMSD between the best template and the native structure are 84% and 2.6 Å, with an average sequence identity of 13% in the aligned regions. These results are compatible with the notion of the completeness of the PDB.

Because *SAL* structural alignments can contain a number of gaps, it might not be possible to build biologically useful models [77], in which case, the conclusion on the completeness of the PDB, while of fundamental interest, would not have practical applications. On the other hand, if the PDB were complete and useful models could be constructed, then, in principle, the protein folding problem could be solved, if one defines the protein folding problem on a purely structural level, i.e. building statistically significant models that have similar topology to native (e.g. with RMSD <6.5 Å). However, to make this conclusion a reality, the development of better threading algorithms to detect all such fold similarities are required. Using the templates and alignments identified from *SAL*, Skolnick et al demonstrated for the 1489 proteins in the *PDB200* benchmark set that: (1). Reasonable full length models can be built by either *MODELLER* [42] or *TASSER*, a newly developed algorithm for threading/assembly/refinement (see below; see also Figure 3 for an schematic overview). Therefore, the conclusion on the completeness of the PDB is of practical interest. (2). Using *TASSER*, consistent improvement of the models from the best structural alignments is demonstrated. (3). Significant improvements in loop modeling are found.

On average, as stated above, from *SAL*, the average RMSD of the structural alignments to native is 2.6 Å with 84% coverage. Skolnick et al applied *TASSER* [122] to the build/refine full-length models for the *PDB200* benchmark set. The *TASSER* final models show improvement over their initial template alignments. Over the same aligned regions, on average, the RMSD is reduced to 1.9 Å. Many low resolution templates improve by refinement to structures with an acceptable resolution for biochemical function annotation [77]. For the entire chain, almost all, but two, targets (with dangling termini involved in intermolecular interactions) have a RMSD < 6 Å for the best of the top five models with an average rank of 1.7 and an average RMSD to native of 2.3 Å. In fact, 97% of the target proteins have a global RMSD < 4 Å. For the rank one cluster (the highest structure density cluster), the average RMSD to native is 2.4 Å. The average RMSD of the best of top five *MODELLER* (a widely used comparative modeling program) models is 3.7 Å, with average rank of 2.9. In general, *TASSER* does a better job in the unaligned regions compared to *MODELLER*, especially for low coverage templates (See Figure 2 below). Looking at those targets with more than 90% coverage (437 in total), the average RMSD of the full length chain models generated by *TASSER* and *MODELLER* are fairly close, i.e. 1.6 Å and 2.2 Å respectively. However, for targets with initial alignment coverage below 75% (386 in total), the average RMSD from native to models by *TASSER* and *MODELLER* are 2.9 Å and 6.1 Å respectively, a significant difference. Overall, in 1120(102) targets, *TASSER* (*MODELLER*) models have lower RMSD to native. In essentially all targets, using the structural alignments provided by *SAL*, reasonable full-length models could be built. Therefore, if one could find the templates and corresponding alignments, given the set of already solved structures in the PDB, these results are highly suggestive that the protein folding problem could be solved for single domain proteins, if one defines the solution as the ability to generate models with a backbone RMSD below 4 Å.

In Figure 2, a detailed comparison of the final models with respect to the template in the aligned regions is plotted. *TASSER* models (Figure 2A & 2B) often show obvious improvement, especially when templates are more than 3 Å away from native. As shown in Figure 2B, for initial template aligned regions with a RMSD from native ranging from 2 to 3 Å, for around 61% of these cases, the models have at least a 0.5 Å improvement; and for targets having initial template aligned regions with a RMSD from native ranging from 3 to 4 Å, for around 49% of these cases, the models have at least 1.0 Å improvement. This improvement occurs because the force field takes consensus information from multiple templates (the top 5 templates are used) [59], as well as the clustering procedure and the energy terms in *TASSER* [122, 123]. Often, the ability to refine models from the best structural alignments (more precisely, using the best structural alignments provided by *SAL*) is demonstrated. In contrast, Figures 2C & 2D, show the comparison between the models generated by *MODELLER* and the initial template alignments. Mainly, *MODELLER* keeps the topology of models near the template [124, 125]. However, sometimes (~10% of cases), the *MODELLER* models are > 1 Å worse than the initial template values.

Here an unaligned or “loop” region is defined as a piece of sequence lacking a coordinate assignment in the *SAL* template alignment. Since no spatial information is provided, modeling the unaligned or loop regions is difficult [125]. Following Fiser et al. [125], two measures of model accuracy are calculated: $\mathrm{RMSD}_{\mathrm{local}}$ denotes the root-mean-square deviation between the native and the modeled loop with direct superposition of the unaligned region and measures the local conformational accuracy. $\mathrm{RMSD}_{\mathrm{global}}$ is the root-mean-square deviation between the native and modeled loop after superposition of up to 5 neighboring stem residues on each side of the loop and measures both the accuracy of the local conformation and its global orientation with respect to the rest of the protein. There are 11,380 unaligned/loop regions ranging from 1 to 84 residues in length in the 1489 targets. In Figures 2E & 2F, the average values of $\mathrm{RMSD}_{\mathrm{local}}$ and $\mathrm{RMSD}_{\mathrm{global}}$ of *TASSER* and *MODELLER* models versus loop length are presented. In both cases, the accuracy decreases with increasing loop size. For all size ranges, *TASSER* models have lower average $\mathrm{RMSD}_{\mathrm{local}}$ and $\mathrm{RMSD}_{\mathrm{global}}$. Focusing on the unaligned loops ≥4 residues in length, there are 1675 cases with an average length of 8.8 residues. *TASSER* shows obviously better control of loop orientations. For example, in 1/3 of the cases, *TASSER* generates models with a $\mathrm{RMSD}_{\mathrm{global}}$<3 Å, while the fraction of *MODELLER* models having a $\mathrm{RMSD}_{\mathrm{global}}$<3 Å is around 1/7. Clearly, while the problem of loop modeling is definitely not solved, some progress is being made.

The "New Fold" targets in CASP5 [126] were also examined by Skolnick et al, since by definition these targets putatively adopt a novel fold never seen in the *PDB* [194]. Using *TASSER*, acceptable models can be built from the initial *SAL* template alignments with an average RMSD from native of 2.87 Å for the first predicted model. Hence, these putative NF targets have templates in the *PDB* that give reasonable structural alignments and full-length models.

THE PROSPECTOR_3 THREADING ALGORITHM

In the past year, Skolnick and coworkers developed an improved threading algorithm, PROSPECTOR_3[50], that is designed to identify analogous as well as homologous templates. The scoring function includes close and distant sequence profiles, secondary structure predictions from *PSIPRED* [28] and a variety of side chain contact pair potentials supplemented by predicted side chain contacts (consensus contacts in at least weakly scoring templates). Alignments are generated using a Needleman-Wunsch global alignment algorithm [51]. Based on score significance, target sequences are classified into three categories: If *PROSPECTOR_3* has at least one significant hit with Z-score, Z, (the energy in standard deviation units relative to mean) above 15 or at least two structurally consistent template hits of Z-score above 7, these targets have high confidence to have a correct template and a good alignment, and the target is assigned to the "*Easy* set". Note that *Easy* does not mean that they are trivially identified; indeed, in the *PDB200* benchmark, *PROSPECTOR_3* correctly assigns more than twice the number of targets to their correct templates (just using the *Easy* set) as *PSI-BLAST* [127]. Sequences that either hit a single template with 7<Z<15 or hit multiple templates lacking a significant consensus structure are assigned to the "*Medium* set"; these have the correct fold identified in most cases, but their alignment may be incorrect. Finally, sequences not assigned to a template belong to the "*Hard* set"; from the point of view of the algorithm, they are New Folds, but based on the finding of the completeness of the *PDB* [16], (almost) all proteins should be assigned to either the *Easy* or *Medium* set by a "perfect" threading algorithm.

PROSPECTOR_3 was applied to the comprehensive *PDB200* benchmark set described above, that are no more than 30% identical to any threading template. In the latest version (somewhat better than published work of Skolnick et al [50], reflecting minor improvements), there are 915 *Easy* protein targets. 791 have a RMSD to native < 6.5 Å. The average contact prediction accuracy is 46%. Continuous aligned regions provide rather accurate (~90% accuracy) native-like fragments that can be used in structure assembly. The level of contact prediction accuracy combined with the fact that continuous fragments are quite accurate motivated the development of *TASSER*, described in the next Section. *67% of the residues obtained from the threading alignments have the same alignment to template as the best structural alignments using SAL*. Threading (structural) alignment refers to the alignment provided by *PROSPECTOR_3* (*SAL*). In addition, 97% of *Easy* targets have a template whose *SAL* structural alignment has a RMSD < 6.5 Å, with an average RMSD of 2.4 Å and 82% average coverage.

PROSPECTOR_3 assigns 565 *Medium* set proteins, 149 with a RMSD< 6.5 Å and 44% average coverage. However, 91% have good *SAL* structural alignments with an average RMSD of 3.8 Å and 50% coverage. The issue is to uncover these alignments (which as seen above can be used to build good models). Combining *Easy/Medium* sets, 65% (94%) of targets have good threading (structural) alignments and the average target/template sequence identity is 22%. Since all targets have good templates in the template library [16], the fact that roughly 1/3 are not identified points out that improvements in *PROSPECTOR_3* are needed. However, consistent with the notion of *PDB* completeness, there are only 19 *Hard* targets.

DEVELOPMENT AND BENCHMARKING OF *TASSER*

Having a set of threading templates, the next thing one wants to do is to build a full length model and to refine the structure so that the regions that have corresponding template alignments move closer to the native state than the template on which they are based. To achieve these two objectives, the Skolnick group has developed the *TASSER*, Threading/ASSEmbly/Refinement algorithm, an overview of which is schematically depicted in Figure 3.

The protein model is described by the alpha-carbon (C_α) atoms and off-lattice side chain centers of mass (SG). The chain is divided into continuous aligned regions extracted from *PROSPECTOR_3* (> 5 residues), whose local conformation is kept essentially unchanged during assembly, and gapped regions that will be treated by *ab initio* methods. The C_αs of these *ab initio* residues lie on an underlying cubic lattice (by discretizing the conformational space, lattices can improve the rate of conformational sampling), while the C_αs of aligned residues are excised from the threading template and are off-lattice (this is done because it is very difficult to move preconstructed fragments around on a lattice. Also lattices introduce an error in the local representation). In a certain sense, *TASSER* represents a convergence of the *ROSETTA*[30] and *TOUCHSTONE II* [128] approaches. However, *ROSETTA* [22] uses small fragments (3~9 residues), and since the conformational search is carried out using large-scale moves (by switching between different local segments), the acceptance rate of *ROSSETTA* movements significantly decreases with increasing fragment size. Here, the threading-based fragments are longer (~20.7 residues on average), the conformational entropy is significantly reduced and more native-like interactions are retained. Movements consist of scaled continuous translations and rotations, allowing for the successful movement of all size substructures. The potential includes predicted secondary structure propensities from *PSIPRED* [129], backbone hydrogen bonds, consensus predicted side chain contacts from *PROSPECTOR_3* [50], statistical short-range correlations and hydrophobic interactions [122]. The combination of energy terms was optimized by maximizing the correlation between the RMSD of decoy structures to native and the energy for 100 nonhomologous training proteins (extrinsic to the *PDB200* benchmark); each with 60,000 decoys. This gave a funnel-like energy landscape, with a correlation coefficient of 0.7 [122] for the training set. For 200 randomly chosen testing proteins in the *PDB200* benchmark set, the correlation coefficient between the energy and RMSD is 0.69; i.e. it is essentially the same for both training and testing proteins.

The next task is to apply the *TASSER* algorithm to a comprehensive benchmark set representative of all the proteins in the PDB below a certain size. The goal here is to have a sufficiently comprehensive set that the results are truly representative. When relatively small sets of proteins are used to test a given algorithm, often the parameters are implicitly optimized so that success is found for the benchmark, but not generally. If say, one considers 100 proteins, and a given variant of a folding algorithm folds 3 additional proteins, does this mean that on average the algorithm is 3% better? In other words, the 3 folded proteins may or may not be representative. However, if benchmarking is done on all representative folds in the PDB, improvements will be statistically significant, and one can ascertain in general what are the strengths and weaknesses of a given algorithm. This will accelerate progress. On the other hand, such large scale benchmarking on thousands of proteins is very CPU intensive, and considerable computational resources are required to carry out the calculations.

APPLICATION TO THE *PDB200* BENCHMARK

With the goal of comprehensive benchmarking, application of *TASSER* to the *PDB200* benchmark set gave the following: There are obvious improvements for almost all quality templates, with the biggest improvement for the poorer quality template alignments (initial RMSD > 8 Å); these mainly belong to the *Medium* and *Hard* sets. For good templates (mostly *Easy* set targets), the alignments are much less gapped, and the tertiary contact restraints from *PROSPECTOR_3* are more consistent. For initial models with a 4~5 Å (2~3 Å) RMSD from native, 58% (43%) of the targets improve by at least 1 (0.5) Å. These results are consistent (see Figure 2) with those when structural alignments are used and show a systematic improvement in model quality. For most initially good templates, (mainly from the *Easy* set) with an initial RMSD of 2~6 Å to native, there is consistently about a 1~3 Å improvement because of the better local structure and side chain group packing following optimization. The final alignments in *MODELLER*[125] tend to be much closer to the initial template alignments. This is not entirely fair since *MODELLER* was designed to fold homologous proteins, and such protein pairs are excluded here.

Turning to loop modeling, considering unaligned/loop regions that have lengths ≥4 residues, the average RMSD by *TASSER* and *MODELLER* are 6.7 Å and 14.9 Å respectively. Using a RMSD cutoff of <4 Å, *MODELLER* gives successful results in 12% of the cases, while *TASSER* is successful in 35% of the cases. These results are slightly worse than when structural alignments are used because of the lower accuracy of the core (See Figure 2E & 2F).

As shown in Figure 4A, defining foldable cases as those where one of the top 5 structures has a RMSD to native below 6.5 Å (a statistically significant value [130], but any reasonable cutoff can be used), the overall success rate for *TASSER* full-length models is 66% (=989/1489). The fraction of targets having an RMSD < 6.5 Å in the aligned regions increases from 65% to 79% after *TASSER* refinement. Furthermore, *TASSER* does not show significant bias to secondary structure class. The success rates for α-, β-, and αβ-proteins are 69%, 61%, and 69% respectively. Nevertheless, a dependence on protein size exists. For targets <120 residues, the success rate is 73%; but for targets >120 residues, it is 58%. All results including threading templates, structure trajectories, and final models for each of the targets are available at http://bioinformatics.buffalo.edu/abinitio/1489.

APPLICATION TO THE *PDB300* BENCHMARK

To explore the ability of *TASSER* to treat larger proteins[122], Skolnick and coworkers examined a second comprehensive *PDB* benchmark set, the *PDB300* set, of 745 proteins ranging in length from 201 to 300 residues; 258 have more than one domain [131]. No pair of target protein sequences has > 35% sequence identity; also, all proteins > 35% identity are excluded from the template library. *PROSPECTOR_3* identifies 593 *Easy* set proteins; 441 have good threading alignments (whose RMSD from native < 6.5 Å), with an average RMSD of 3.6 Å, 83% coverage and 21% sequence identity to their templates. There are 150 *Medium* and 2 *Hard* targets. Using this information, Figure 4B shows the *TASSER* results for the percent of predicted targets with a given RMSD, with single and multiple domain protein targets presented separately. The success rate for all *PDB300* targets is 55%. 61% of single domain proteins have the best of top five models with a RMSD to native < 6.5 Å. This is slightly less than the success rate of 66% for single domain proteins ≤ 200 residues [122]. For multiple domain proteins, 43% have a RMSD < 6.5 Å for the best of top 5 models. But 2/3 of these multiple domain targets have at least one domain (average length of 144 residues) with a RMSD<6.5 Å. Thus, domains are often correctly predicted, but not their mutual orientation. This is a significant problem that must be addressed.

Similar to the case of proteins ≤ 200 residues, *TASSER* gives significant improvements with respect to the initial alignments. For example, for initial alignments with a RMSD between 4-5 Å, in 53% of the cases, the final models improve by at least 1 Å. Turning to loop modeling and focusing on unaligned/loops ≥4 residues, there are in total 1809 cases with average length 12.2 residues. In around 1/3 of cases, the *TASSER* loop modeling procedure has acceptable accuracy.

RESULTS FOR TRANSMEMBRANE PROTEINS

There are 18 large membrane proteins in the *PDB300* benchmark set. For 1/3, *TASSER* generates at least one model in the top five that has a RMSD to native below 5.5 Å. For the *PDB200* benchmark (proteins 41-200 residues), there are 20 membrane proteins, with a success rate of 45%. Among the total of 15 foldable membrane targets in both sets, for 10, *PROSPECTOR_3* hits at least one other nonhomologous transmembrane template; in the remaining five, *PROSPECTOR_3* hit globular proteins with regular helical structures consistent with the target structures, which provided the opportunity for *TASSER* to assemble/refine the models. Figure 5, shows three typical results for membrane proteins: 1jgjA, 1fqyA, and 1bh3_, with the well-known GPCR rhodopsin, 1jgjA having the highest resolution. Their best template hits by *PROSPECTOR_3* are respectively: 1ap9_ (1.47 Å over 96% coverage and 29% sequence identity), 1fx8A (5.20 Å over 92% coverage and 29% sequence identity), and 2por_(13.44 Å over 88% coverage and 22% sequence identity). The final models have a RMSD to native of 1.1/0.89 Å, 3.3/3.1 Å, and 5.3/5.2 Å over the full-length/aligned-regions respectively. This shows that *TASSER* improves threading alignments and builds reasonable loops for membrane proteins.

COMPARISON OF *TASSER* MODELS WITH NMR STRUCTURES

For all representative proteins ≤300 residues (in both the *PDB200* and *PDB300* benchmark sets) that have corresponding multiple NMR structures in the PDB, ≈20% of the models generated by *TASSER* are closer to the NMR structure centroid than the farthest individual NMR model. Note that no experimental information is employed in this set of predictions. Some representative examples for proteins belonging to each of the three secondary structure classes are shown in Figure 6. While this represents encouraging progress; nevertheless there remain the 80% of proteins with NMR structures that are not predicted at the level of experimental resolution. These remain an outstanding challenge.

EXTENSION OF THREADING TO PREDICT QUATERNARY STRUCTURE

Over the past several years, multimeric threading algorithm, MULTIPROSPECTOR was developed and benchmarked by Skolnick and coworkers [119]. The approach consists of two phases: First, traditional single threading is applied to generate a set of candidate structures. Then, for those proteins whose template structures are part of a known complex, they rethread both partners and now include a protein-protein interfacial energy. A database of multimeric protein template structures was constructed [118], interfacial pairwise potentials derived, and empirical indicators to identify dimers based on their threading Z-score and the magnitude of the interfacial energy was established. The authors tested the algorithm on a benchmark set comprised of 58 homodimers, 20 heterodimers, and 96 monomers scanned against 3900 representative template structures. The method correctly recognized and assigned 54 homodimers, all 20 heterodimers, and 91 monomers, and satisfactory performance was demonstrated [119].

APPLICATION TO PROTEOMES

PROSPECTOR_3 RESULTS

To examine the generality of the *PDB200* benchmark results, Skolnick and coworkers applied *PROSPECTOR_3* to ORFS ≤ 200 residues in the *E. coli* [132], *M. genitalium* [133], and *S. cerevisiae*[134] proteomes. Unlike the benchmark, here homologous proteins are allowed. An overview is presented with details given elsewhere; see http://www.bioinformatics.buffalo.edu/resources/genomethreading/. For *E. coli* [132], there are 1360 ORFs ≤ 200 residues. *PROSPECTOR_3* assigns 61% to the *Easy* set (82% average coverage) and 38% to the *Medium* set (51% average coverage). In contrast, Peitsch et al [135] produced assignments for ~10-15% of the entire proteome. Using *PSI-BLAST* [127], Hegyi et al. [136] assigned 28% of all *E. coli* ORFs to *SCOP* domains. In *PEDANT* [40], 31% of *E. coli* ORFs ≤ 200 residues have a *PSI-BLAST* hit to *PDB* structures. In *GTOP* [137], *Reverse PSI-BLAST* [138] assigned 35% of *E. coli* ORFs ≤ 200 residues to PDB structures. The *M. genitalium* [133] proteome has 128 ORFs ≤ 200 residues. *PROSPECTOR_3* assigns 73% to the *Easy* set (87% average coverage), and 27% to the *Medium* set (54% average coverage). In *S. cerevisiae*[139], there are 1496 ORFs ≤ 200 residues. *PROSPECTOR_3* assign 53% to the *Easy* set (75% average coverage) and 45% to the *Medium* set (65% average coverage). There are few putative New Folds ORFs in all three proteomes.

TASSER RESULTS

TASSER was also applied to all ORFs in the *E. coli* proteome [132] ≤ 200 residues. Based on the *PDB* benchmarks, a confidence, C-score, is defined that is a function of cluster density, the RMSD of cluster members from the cluster centroid and the threading template Z-score (see eq. 1 of ref [122]). For the same C-score cutoff in the *PDB200* benchmark that gives a false positive/negative rate of 12.4%/14.7%, 68% of *E. coli* ORFs should have acceptable predictions. All results are available at http://www.bioinformatics.buffalo.edu/genome/ecoli. According to *MEMSAT* [140] ~23% of these *E. coli* ORFs have transmembrane regions. All *TASSER* predicted first rank models have at least one long (putative transmembrane) helix consistent with *MEMSAT*. Using the C-score, 47% of the ORFs have >80% probability for models with a RMSD < 6.5 Å. Furthermore, signal peptides are not masked out, and 149 ORFs have annotated signal peptides in *SWISS-PROT* [141]. Due to their composition, *PROSPECTOR_3* does not align the majority of signal peptide residues, and due to the resulting lack of predicted contacts, these peptides lie outside the predicted compact core. A possibility to be pursued is to use this method to identify signal sequences.

APPLICATION OF *MULTIPROSPECTOR* TO *S. CEREVISIAE*

Using *MULTIPROSPECTOR*, each possible pair of interactions among more than six thousand encoded proteins is evaluated against a dimer database of 768 complex structures by using a confidence estimate of the fold assignment and the magnitude of the interfacial potentials. 7,321 interactions are predicted involving 1,256 proteins. After filtering by subcellular colocalization, there are 2,028 heterodimer interactions. From mRNA abundance analysis, the *MULTIPROSPECTOR* method does not bias towards high abundance proteins. The predicted interactions are then compared to other large-scale methods and to high confidence interactions defined as those supported by two or more other methods [18]. 374 of the predictions are found by at least one other study, comparable to the overlap between two other methods. Based on functional category assignment, *MULTIPROSPECTOR* predictions have a similar distribution as high confidence interactions.

CONCLUSION

At this juncture, it is apparent that considerable progress is being made in the field of protein structure prediction, with the greatest success seen for knowledge-based approaches that extend comparative modeling and threading. At present, based on very large scale benchmarking, for weakly/nonhomologous proteins, one can expect to produce low-resolution structures for about 2/3 of all proteins. Given the observation that the PDB is complete for low-to-moderate resolution single domain proteins, the outstanding challenge is to develop methods to identify the roughly 1/3 of proteins that cannot be recognized by contemporary approaches. Furthermore, progress is being made on generating predictions where the model is closer to the native structure than to the template on which it is based. Part of the reason for the recent relative success is the comprehensive testing on all representative PDB structures so that one can identify both the strengths and weaknesses of a given approach. In the past, relatively small scale benchmarking was done, where it was difficult to establish the generality of the conclusions. Another reason is the improved correlation of energy and structure quality. This is not to say that existing potentials are perfect, for certainly they are not, but rather that procedures to derive better potentials are starting to bear fruit.

There remain a number of outstanding problems that must be addressed: With regards to low-resolution modeling, existing approaches to predict the relative orientation of multiple domain proteins often fail when the domains adopt a different orientation from the template. This is same issue as the inability to predict good global orientations for long loops even when (as is often the case) their internal conformation is well predicted. This reflects problems with the force field. Similarly, it is still not possible in general to refine the low-resolution structures to higher quality structures at atomic detail. Whether this is an issue of conformational sampling or problems with existing atomic force fields or both remains to be established. In that regard, Skolnick and coworkers have embarked on a similar large scale benchmarking effort to identify the outstanding unresolved issues with the goal of making progress in detailed atomic model refinement. Indeed, one would like to supercede the current generation of knowledge-based approaches with more fundamental physics based approaches. The next issue that must be addressed is the prediction of protein-protein interactions and the quaternary structure of the resulting complexes. Here, the field of structure prediction is in its infancy; approaches similar to *ROSETTA* and *TASSER* generalized to multimers represent promising avenues of investigation. At the end of the day, one goal of protein structure prediction is to provide models that are of sufficient quality that they can provide functional insights. While much remains to be done, there is now cause for optimism that the progress is being made to achieve this objective.

ACKNOWLEDGMENTS

This research was supported in part by NIH grants GM-37408, GM-48835, and RR-12255. Stimulating discussions and essential contributions of our colleagues Drs. A. Arakaki, D. Kihara, L. Lu and H. Lu are gratefully acknowledged.

References

[1] Venter, J.C., et al., The sequence of the human genome. *Science*, 2001. 291(5507): 1304-1351, 2001.

[2] Wiley, S.R., Genomics in the real world. *Curr Pharm Des*, 1998. 4(5): 417-422, 1998.

[3] Betz, S.F., S.M. Baxter, and J.S. Fetrow, Function first: a powerful approach to post-genomic drug discovery. *Drug Discov Today*, 2002. 7(16): 865-871, 2002.

[4] Pearson, W.R., Effective protein sequence comparison. *Methods Enzymol*, 1996. 266227-258, 1996.

[5] Kinch, L.N., et al., CASP5 assessment of fold recognition target predictions. *Proteins*, 2003. 53 Suppl 6395-409, 2003.

[6] Wallace, A.C., R.A. Laskowski, and J.M. Thornton, Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci*, 1996. 5(6): 1001-1013, 1996.

[7] Kleywegt, G.J., Recognition of spatial motifs in protein structures. *J Mol Biol*, 1999. 285(4): 1887-1897, 1999.

[8] Skolnick, J., J.S. Fetrow, and A. Kolinski, Structural genomics and its importance for gene function analysis. *Nat Biotechnol*, 2000. 18(3): 283-287, 2000.

[9] Baker, D. and A. Sali, Protein structure prediction and structural genomics. *Science*, 2001. 294(5540): 93-96, 2001.

[10] Aloy, P., et al., Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol*, 2001. 311(2): 395-408, 2001.

[11] Turcotte, M., S.H. Muggleton, and M.J. Sternberg, Automated discovery of structural signatures of protein fold and function. *J Mol Biol*, 2001. 306(3): 591-605, 2001.

[12] Gerstein, M., et al., Structural genomics: current progress. *Science*, 2003. 299(5613): 1663, 2003.

[13] Vitkup, D., et al., Completeness in structural genomics. *Nat Struct Biol*, 2001. 8(6): 559-566, 2001.

[14] Moult, J. and E. Melamud, From fold to function. *Curr Opin Struct Biol*, 2000. 10(3): 384-389, 2000.

[15] McGuffin, L.J. and D.T. Jones, Targeting novel folds for structural genomics. *Proteins*, 2002. 48(1): 44-52, 2002.

[16] Kihara, D. and J. Skolnick, The PDB is a Covering Set of Small Protein Structures. *J Mol Biol*, 2003. 334(4): 793-802, 2003.

[17] Alberts, B., et al., *Molecular biology of the cell*. 3rd ed. 1994, New York: Garland Pub. xliii, 1294, [1267].

[18] Mering, C.V., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., & Bork, P., Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 20021-5, 2002.

[19] Marti-Renom, M.A., et al., Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 2000. 29291-325, 2000.

[20] Bowie, J.U., R. Luthy, and D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 1991. 253164-170, 1991.

[21] Liwo, A., et al., Protein structure prediction by global optimization of a potential energy function. *Proc Natl Acad Sci U S A*, 1999. 96(10): 5482-5485, 1999.

[22] Simons, K.T., C. Strauss, and D. Baker, Prospects for *ab initio* protein structural genomics. *J. Mol. Biol.*, 2001. 3061191-1199, 2001.

[23] Kihara, D., et al., TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci U S A*, 2001. 98(18): 10125-10130, 2001.

[24] Westbrook, J., et al., The Protein Data Bank: unifying the archive. *Nucleic Acids Res*, 2002. 30(1): 245-248, 2002.

[25] Kopp, J. and T. Schwede, The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res*, 2004. 32 Database issueD230-234, 2004.

[26] John, B. and A. Sali, Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res*, 2003. 31(14): 3982-3992, 2003.

[27] Xu, D., et al., Application of PROSPECT in CASP4: characterizing protein structures with new folds. *Proteins*, 2001. Suppl 5140-148, 2001.

[28] McGuffin, L.J., K. Bryson, and D.T. Jones, The PSIPRED protein structure prediction server. *Bioinformatics*, 2000. 16(4): 404-405, 2000.

[29] Skolnick, J., et al., A unified approach to protein structure prediction. *Proteins*, 2003. CASP5 Suppl469-479, 2003.

[30] Bradley, P., et al., Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins*, 2003. 53 Suppl 6457-468, 2003.

[31] Aloy, P., et al., Predictions without templates: new folds, secondary structure, and contacts in CASP5. *Proteins*, 2003. 53 Suppl 6436-456, 2003.

[32] Liwo, A., et al., A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: application to the UNRES force field. *Proc Natl Acad Sci U S A*, 2002. 99(4): 1937-1942, 2002.

[33] Venclovas, C., et al., Assessment of progress over the CASP experiments. *Proteins*, 2003. 53 Suppl 6585-595, 2003.

[34] Tramontano, A. and V. Morea, Assessment of homology-based predictions in CASP5. *Proteins*, 2003. 53 Suppl 6352-368, 2003.

[35] Iliopoulos, I., et al., Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics*, 2003. 19(6): 717-726, 2003.

[36] de Bakker, P.I., et al., Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. *Proteins*, 2003. 51(1): 21-40, 2003.

[37] Bonneau, R., et al., Functional inferences from blind ab initio protein structure predictions. *J Struct Biol*, 2001. 134(2-3): 186-190, 2001.

[38] Skolnick, J. and J.S. Fetrow, From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends Biotechnol*, 2000. 18(1): 34-39, 2000.

[39] Gerstein, M., Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins*, 1998. 33(4): 518-534, 1998.

[40] Frishman, D., et al., The PEDANT genome database. *Nucleic Acids Res*, 2003. 31(1): 207-211, 2003.

[41] Yamaguchi, A., et al., Enlarged FAMSBASE: protein 3D structure models of genome sequences for 41 species. *Nucleic Acids Res*, 2003. 31(1): 463-468, 2003.

[42] Pieper, U., et al., MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res*, 2004. 32 Database issueD217-222, 2004.

[43] Maiorov, V.N. and G.M. Crippen, Contact potential that recognizes the correct folding of globular proteins. ***J. Mol. Biol.***, 1992. 277876-888, 1992.

[44] Sippl, M.J., & Weitckus, S., Detection of Native-Like Models for Amino Acid Sequences of Unknown Three-Dimensional Structure in a Database of Known Protein Conformations. *Proteins*, 1992. 13258-271, 1992.

[45] Skolnick, J. and D. Kihara, Defrosting the frozen approximation: PROSPECTOR--a new approach to threading. *Proteins*, 2001. 42(3): 319-331, 2001.

[46] Bryant, S.H. and C.E. Lawrence, An empirical energy function for threading protein sequence through the folding motif. *Proteins*, 1993. 16(1): 92-112, 1993.

[47] Godzik, A., Skolnick, J. & Kolinski, A., A Topology Fingerprint Approach to the Inverse Folding Problem. ***J. Mol. Biol.***, 1992. 227227-238, 1992.

[48] McGuffin, L.J. and D.T. Jones, Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, 2003. 19(7): 874-881, 2003.

[49] Zhang, B., et al., Similarities and differences between nonhomologous proteins with similar folds: evaluation of threading strategies. *Fold Des*, 1997. 2(5): 307-317, 1997.

[50] Skolnick, J., D. Kihara, and Y. Zhang, Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins*, 2004. 56(3): 502-518, 2004.

[51] Needleman, S.B. and C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 1970. 48(3): 443-453, 1970.

[52] Orengo, C.A. and W.R. Taylor, A local alignment method for protein structure motifs. *J Mol Biol*, 1993. 233(3): 488-497, 1993.

[53] Panchenko, A.R., A. Marchler-Bauer, and B.S. H., Combination of threading potentials and sequence profiles improves fold recognition. ***J. Mol. Biol.***, 2000. 2961319-1331, 2000.

[54] Lathrop, R.H., An anytime local-to-global optimization algorithm for protein threading in theta (m2n2) space. *J Comput Biol*, 1999. 6(3-4): 405-418, 1999.

[55] Gerstein, M. and M. Levitt, Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *Proc Int Conf Intell Syst Mol Biol*, 1996. 459-67, 1996.

[56] Shindyalov, I.N. and P.E. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, 1998. 11(9): 739-747, 1998.

[57] Kosinski, J., et al., A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins*, 2003. 53 Suppl 6369-379, 2003.

[58] Kinch, L.N., et al., CASP5 target classification. *Proteins*, 2003. 53 Suppl 6340-351, 2003.

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

[60] Wallner, B., H. Fang, and A. Elofsson, Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. *Proteins*, 2003. 53 Suppl 6534-541, 2003.

[61] Chivian, D., et al., Automated prediction of CASP-5 structures using the Robetta server. *Proteins*, 2003. 53 Suppl 6524-533, 2003.

[62] Eyrich, V.A., et al., CAFASP3 in the spotlight of EVA. *Proteins*, 2003. 53 Suppl 6548-560, 2003.

[63] Rychlewski, L., D. Fischer, and A. Elofsson, LiveBench-6: large-scale automated evaluation of protein structure prediction servers. *Proteins*, 2003. 53 Suppl 6542-547, 2003.

[64] Holm, L. and C. Sander, Touring protein fold space with Dali/FSSP. *Nucleic Acids Res*, 1998. 26(1): 316-319, 1998.

[65] Grindley, H.M., et al., Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J Mol Biol*, 1993. 229(3): 707-721, 1993.

[66] Mizuguchi, K. and N. Go, Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Eng*, 1995. 8(4): 353-362, 1995.

[67] Bachar, O., et al., A computer vision based technique for 3-D sequence-independent structural comparison of proteins. *Protein Eng*, 1993. 6(3): 279-288, 1993.

[68] An, Y. and R.A. Friesner, A novel fold recognition method using composite predicted secondary structures. *Proteins*, 2002. 48(2): 352-366, 2002.

[69] Kedem, K., L.P. Chew, and R. Elber, Unit-vector RMS (URMS) as a tool to analyze molecular dynamics trajectories. *Proteins*, 1999. 37(4): 554-564, 1999.

[70] Ortiz, A.R., C.E. Strauss, and O. Olmea, MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci*, 2002. 11(11): 2606-2621, 2002.

[71] Orengo, C.A., et al., CATH--a hierarchic classification of protein domain structures. *Structure*, 1997. 5(8): 1093-1108, 1997.

[72] Shindyalov, I.N. and P.E. Bourne, An alternative view of protein fold space. *Proteins*, 2000. 38(3): 247-260, 2000.

[73] Boutonnet, N.S., A.V. Kajava, and M.J. Rooman, Structural classification of alphabetabeta and betabetaalpha supersecondary structure units in proteins. *Proteins*, 1998. 30(2): 193-212, 1998.

[74] Harrison, A., et al., Quantifying the similarities within fold space. *J Mol Biol*, 2002. 323(5): 909-926, 2002.

[75] Yang, A.S. and B. Honig, An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J Mol Biol*, 2000. 301(3): 665-678, 2000.

[76] Lo Conte, L., et al., SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res*, 2002. 30(1): 264-267, 2002.

[77] Arakaki, A.K., Y. Zhang, and J. Skolnick, Large scale assessment of the utility of low resolution protein structures for biochemical function assignment. *Bioinformatics*, 2004. 201087-1096, 2004.

[78] Claudel-Renard, C., et al., Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res*, 2003. 31(22): 6633-6639, 2003.

[79] Henikoff, J.G., et al., Blocks-based methods for detecting protein homology. *Electrophoresis*, 2000. 21(9): 1700-1706, 2000.

[80] Attwood, T.K., et al., PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res*, 2003. 31(1): 400-402, 2003.

[81] Fetrow, J.S., et al., Genomic-scale comparison of sequence- and structure-based methods of function prediction: does structure provide additional insight? *Protein Sci*, 2001. 10(5): 1005-1014, 2001.

[82] Hulo, N., et al., Recent improvements to the PROSITE database. *Nucleic Acids Res*, 2004. 32 Database issueD134-137, 2004.

[83] Hegyi, H. and M. Gerstein, The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol*, 1999. 288(1): 147-164, 1999.

[84] Kihara, D. and J. Skolnick, Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. *Proteins*, 2004. 55(2): 464-473, 2004.

[85] Wallace, A.C., N. Borkakoti, and J.M. Thornton, TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci*, 1997. 6(11): 2308-2323, 1997.

[86] Russell, R.B., Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J Mol Biol*, 1998. 279(5): 1211-1227, 1998.

[87] Fetrow, J.S. and J. Skolnick, Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J Mol Biol*, 1998. 281(5): 949-968, 1998.

[88] Zhao, S., et al., Recognition templates for predicting adenylate-binding sites in proteins. *J Mol Biol*, 2001. 314(5): 1245-1255, 2001.

[89] Hamelryck, T., Efficient identification of side-chain patterns using a multidimensional index tree. *Proteins*, 2003. 51(1): 96-108, 2003.

[90] Liang, M.P., D.L. Brutlag, and R.B. Altman, Automated construction of structural motifs for predicting functional sites on protein structures. *Pac Symp Biocomput*, 2003204-215, 2003.

[91] Peters, K.P., J. Fauck, and C. Frommel, The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J Mol Biol*, 1996. 256(1): 201-213, 1996.

[92] Wei, L., E.S. Huang, and R.B. Altman, Are predicted structures good enough to preserve functional sites? *Structure Fold Des*, 1999. 7(6): 643-650, 1999.

[93] Stark, A. and R.B. Russell, Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res*, 2003. 31(13): 3341-3344, 2003.

[94] Jones, D.T. and L.J. McGuffin, Assembling novel protein folds from super-secondary structural fragments. *Proteins*, 2003. 53 Suppl 6480-485, 2003.

[95] Schmitt, S., D. Kuhn, and G. Klebe, A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol*, 2002. 323(2): 387-406, 2002.

[96] Adams, M.D., et al., The genome sequence of Drosophila melanogaster. *Science*, 2000. 287(5461): 2185-2195, 2000.

[97] Bonneau, R., et al., De novo prediction of three-dimensional structures for major protein families. *J Mol Biol*, 2002. 322(1): 65-78, 2002.

[98] Fetrow, J.S., A. Godzik, and J. Skolnick, Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J Mol Biol*, 1998. 282(4): 703-711, 1998.

[99] Legrain, P., J. Wojcik, and J.M. Gauthier, Protein--protein interaction maps: a lead towards cellular functions. *Trends Genet*, 2001. 17(6): 346-352, 2001.

[100] Fields, S. and O. Song, A novel genetic system to detect protein-protein interactions. *Nature*, 1989. 340(6230): 245-246, 1989.

[101] Sobott, F. and C.V. Robinson, Protein complexes gain momentum. *Curr Opin Struct Biol*, 2002. 12(6): 729-734, 2002.

[102] Uetz, P., et al., A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 2000. 403(6770): 623-627, 2000.

[103] Ito, T., et al., A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 2001. 98(8): 4569-4574, 2001.

[104] Janin, J. and B. Seraphin, Genome-wide studies of protein-protein interaction. *Curr Opin Struct Biol*, 2003. 13(3): 383-388, 2003.

[105] Valencia, A. and F. Pazos, Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol*, 2002. 12(3): 368-373, 2002.

[106] Huynen, M.A., et al., Function prediction and protein networks. *Curr Opin Cell Biol*, 2003. 15(2): 191-198, 2003.

[107] Marcotte, E.M., et al., Detecting protein function and protein-protein interactions from genome sequences. *Science*, 1999. 285(5428): 751-753, 1999.

[108] Enright, A.J., et al., Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 1999. 402(6757): 86-90, 1999.

[109] Overbeek, R., et al., The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, 1999. 96(6): 2896-2901, 1999.

[110] Pazos, F., et al., Correlated mutations contain information about protein-protein interaction. *J Mol Biol*, 1997. 271(4): 511-523, 1997.

[111] Pazos, F. and A. Valencia, Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 2001. 14(9): 609-614, 2001.

[112] Valencia, A. and F. Pazos, *In Structural Bioinformatics*. 2003: John Wiley & Sons.

[113] Tatusov, R.L., et al., The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 2003. 4(1): 41, 2003.

[114] Comet, J.P. and J. Henry, Pairwise sequence alignment using a PROSITE pattern-derived similarity score. *Comput Chem*, 2002. 26(5): 421-436, 2002.

[115] Aloy, P. and R.B. Russell, The third dimension for protein interactions and complexes. *Trends Biochem Sci*, 2002. 27(12): 633-638, 2002.

[116] Adams, J., The proteasome: structure, function, and role in the cell. *Cancer Treat Rev*, 2003. 29 Suppl 13-9, 2003.

[117] Fariselli, P., et al., Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins*, 2001. Suppl 5157-162, 2001.

[118] Lu, L., et al., Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the Saccharomyces cerevisiae proteome. *Genome Res*, 2003. 13(6A): 1146-1154, 2003.

[119] Lu, L., H. Lu, and J. Skolnick, MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, 2002. 49(3): 350-364, 2002.

[120] Henikoff, S. and J.G. Henikoff, Performance evaluation of amino acid substitution matrices. *Proteins*, 1993. 17(1): 49-61, 1993.

[121] Zhang, Y. and J. Skolnick, The protein structure prediction problem could be solved using the current PDB library. *Proc Natl Acad Sci U S A*, 2005. 102: 1029-1034, 2005.

[122] Zhang, Y. and J. Skolnick, Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci U S A*, 2004. 101: 7594-7599, 2004.

[123] Zhang, Y. and J. Skolnick, SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem*, 2004. 25(6): 865-871, 2004.

[124] Sali, A. and T.L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 1993. 234(3): 779-815, 1993.

[125] Fiser, A., R.K. Do, and A. Sali, Modeling of loops in protein structures. *Protein Sci*, 2000. 9(9): 1753-1773, 2000.

[126] Moult, J., et al., Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*, 2003. 53 Suppl 6334-339, 2003.

[127] Altschul, S.F., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 1997. 25(17): 3389-3402, 1997.

[128] Zhang, Y., A. Kolinski, and J. Skolnick, TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys J*, 2003. 85(2): 1145-1164, 2003.

[129] Jones, D.T., Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 1999. 292195-202, 1999.

[130] Reva, B.A., A.V. Finkelstein, and J. Skolnick, What is the probability of a chance prediction of a protein structure with an rmsd of 6 A? *Fold Des*, 1998. 3(2): 141-147, 1998.

[131] Guo, J.T., et al., Improving the performance of DomainParser for structural domain partition using neural network. *Nucleic Acids Res*, 2003. 31944-952, 2003.

[132] Blattner, F.R., etc, The Complete Genome Sequence of *Escherichia coli* K-12. *Science*, 1997. 277(5331): 1453-1474, 1997.

[133] Fraser, C.M., et al., The minimal gene complement of Mycoplasma genitalium. *Science*, 1995. 270(5235): 397-403, 1995.

[134] Mewes, H.M., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., Stocker, S. & Weil, B., MIPS: a Database for Genomes and Protein Sequences. *Nucleic Acids Res.*, 2000. 28(1): 37-40, 2000.

[135] Peitsch, M.C., et al., Large-scale protein modelling and integration with the SWISS-PROT and SWISS-2DPAGE databases: the example of Escherichia coli. *Electrophoresis*, 1997. 18(3-4): 498-501, 1997.

[136] Hegyi, H., et al., Structural genomics analysis: characteristics of atypical, common, and horizontally transferred folds. *Proteins*, 2002. 47(2): 126-141, 2002.

[137] Kawabata, T., et al., GTOP: a database of protein structures predicted from genome sequences. *Nucleic Acids Res*, 2002. 30(1): 294-298, 2002.

[138] Marchler-Bauer, A., et al., CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res*, 2002. 30(1): 281-283, 2002.

[139] Mewes, H.W., et al., MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 2000. 28(1): 37-40, 2000.

[140] Jones, D.T., W.R. Taylor, and J.M. Thornton, A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, 1994. 33(10): 3038-3049, 1994.

[141] Bairoch, A. and R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res*, 1998. 26(1): 38-42, 1998.

[142] Zhang, Y., D. Kihara, and J. Skolnick, Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins*, 2002. 48(2): 192-201, 2002.

Figure 1. Schematic overview of the methodologies employed in Comparative Modeling/Threading**,** and *ab initio* folding.

Figure 2 **A.** Scatter plot of RMSD from native to the final models built by *TASSER* refinements versus RMSD to native in the best initial template alignments identified by *SAL*. The same aligned regions are used in both RMSD calculations. **B.** Using *TASSER*, the fraction of targets with a RMSD improvement *d* greater than some threshold value. Here *d*="RMSD of template"-"RMSD of final model", where each RMSD is calculated over the aligned regions. Each point is calculated with a bin width of 1 Å. **C.** Similar data as in A, but the models are from *MODELLER* refinements. **D.** Similar data as in B, but the models are from *MODELLER* refinements. **E**. $\text{RMSD}_{\text{local}}$ and **F.** $\text{RMSD}_{\text{global}}$ of unaligned/loop regions as a function of loop length. *TASSER* and *MODELLER* models are denoted by triangles and circles respectively. The lines connecting the points serve to guide the eye. The dashed line in F denotes a $\text{RMSD}_{\text{global}}$ cutoff of 7 Å.

Figure 3**.** Overview of the *TASSER* structure prediction methodology that consists of template identification by *PROSPECTOR_3* [50] that provides template fragments and predicted contact restraints, fragment assembly using Parallel Hyperbolic Sampling [142], and fold selection by *SPICKER* clustering [122]. The entire process for 1ayyD is shown.

Figure 4A. For the *PDB*200 benchmark set of proteins, histograms of foldable proteins using *MODELLER* [124] and *TASSER* based on the same templates and alignments from *PROSPECTOR_3* [50]. **B.** For proteins in the *PDB300* benchmark set, using *TASSER,* the histogram of the percent of predicted targets as function of global RMSD, divided into single and multiple domain categories.

Figure 5. Three representative examples of the successful structure prediction of transmembrane proteins by *TASSER*. The thin (thick) lines denote the C_α-backbone of the experimental (predicted) structure. Blue to red runs from the N- to C-terminus. Below the structures are their *PDB* id, the RMSD between the model and native structure, and the length of the protein.

Figure 6. Three representative examples of *TASSER* predicted models that are structurally closer to the NMR structure centroid than some of individual NMR structures. The thick backbone shows the rank-one models predicted by *TASSER*; the wire frame presents the structures satisfying the NMR distance constraints equally well. Blue to red runs from the N- to C-terminus. The RMSD of *TASSER* models to the NMR centroid for 1adr_ (α-protein), 2fnbA (β-protein), and 1dbyA (αβ-protein) are 1.6 Å, 1.9 Å, and 1.1 Å respectively; the maximum RMSD of NMR models to the centroid are 3.6 Å, 2.3 Å, and 1.3 Å, respectively.

Figure 1.

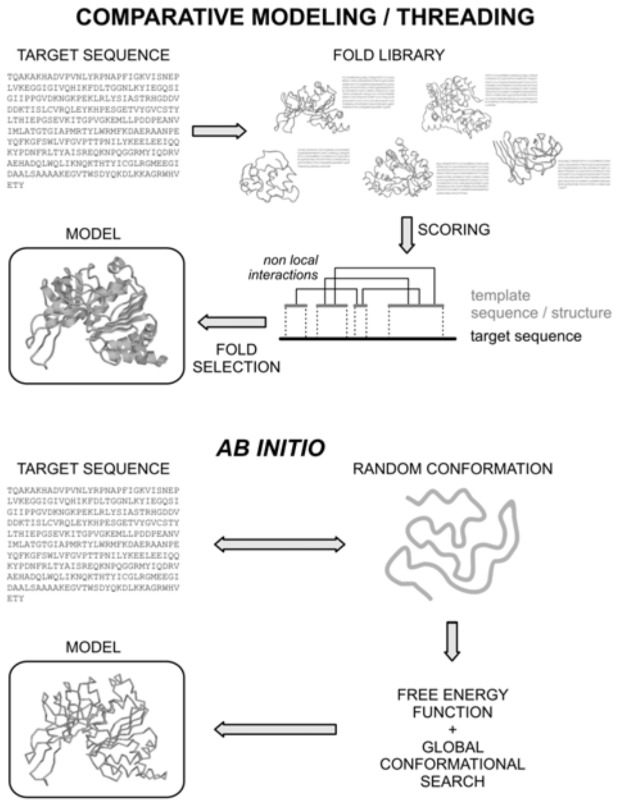

COMPARATIVE MODELING / THREADING
TARGET SEQUENCE
TQAKAKHADVPVNLYRPNAPFIGKVISNEP
LVKEGGIGIVQHIKFDLTGGNLKYIEGQSI
GIIPPGVDKNGKPEKLRLYSIASTRHGDDV
DDKTISLCVRQLEYKHPESGETVYGVCSTY
LTHIEPGSEVKITGPVGKEMLLPDDPEANV
IMLATGTGIAPMRTYLWRMFKDAERAANPE
YQFKGFSWLVFGVPTTPNILYKEELEEIQQ
KYPDNFRLTYAISREQKNPQGGRMYIQDRV
AEHADQLWQLIKNQKTHTYICGLRGMEEGI
DAALSAAAAKEGVTWSDYQKDLKKAGRWHV
ETY
FOLD LIBRARY
SCORING
MODEL
non local
interactions
template
sequence / structure
target sequence
FOLD
SELECTION
AB INITIO
TARGET SEQUENCE
TQAKAKHADVPVNLYRPNAPFIGKVISNEP
LVKEGGIGIVQHIKFDLTGGNLKYIEGQSI
GIIPPGVDKNGKPEKLRLYSIASTRHGDDV
DDKTISLCVRQLEYKHPESGETVYGVCSTY
LTHIEPGSEVKITGPVGKEMLLPDDPEANV
IMLATGTGIAPMRTYLWRMFKDAERAANPE
YQFKGFSWLVFGVPTTPNILYKEELEEIQQ
KYPDNFRLTYAISREQKNPQGGRMYIQDRV
AEHADQLWQLIKNQKTHTYICGLRGMEEGI
DAALSAAAAKEGVTWSDYQKDLKKAGRWHV
ETY
RANDOM CONFORMATION
MODEL
FREE ENERGY
FUNCTION
+
GLOBAL
CONFORMATIONAL
SEARCH

Figure 2.

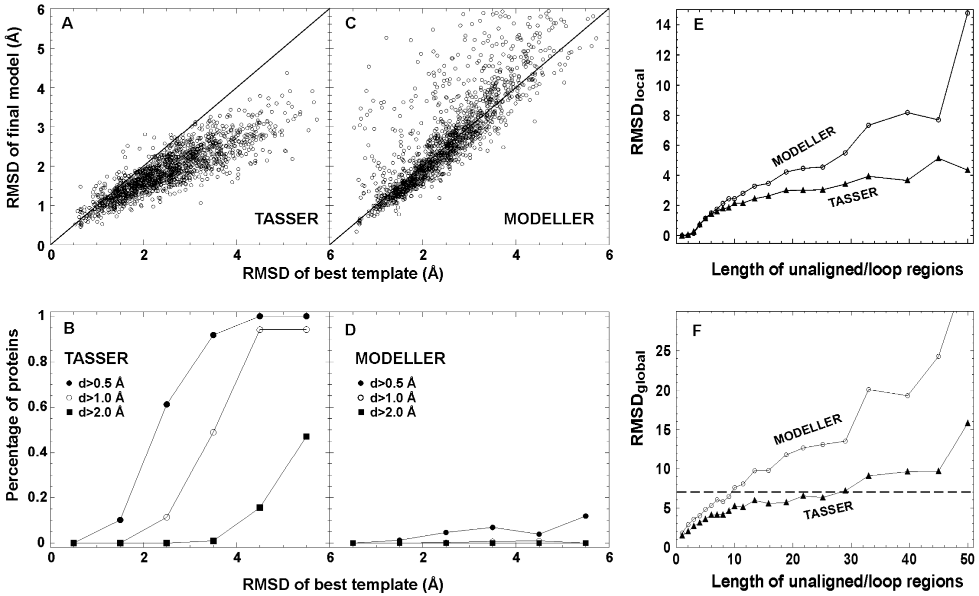

A
TASSER
C
MODELLER
RMSD of final model (Å)
RMSD of best template (Å)
B
TASSER
● d>0.5 Å
○ d>1.0 Å
■ d>2.0 Å
D
MODELLER
● d>0.5 Å
○ d>1.0 Å
■ d>2.0 Å
Percentage of proteins
RMSD of best template (Å)
0
0.2
0.4
0.6
0.8
1
0
1
2
3
4
5
6
0
2
4
6
E
MODELLER
TASSER
$RMSD_{local}$
0
2
4
6
8
10
12
14
Length of unaligned/loop regions
F
MODELLER
TASSER
$RMSD_{global}$
0
5
10
15
20
25
0
10
20
30
40
50
Length of unaligned/loop regions

Figure 3

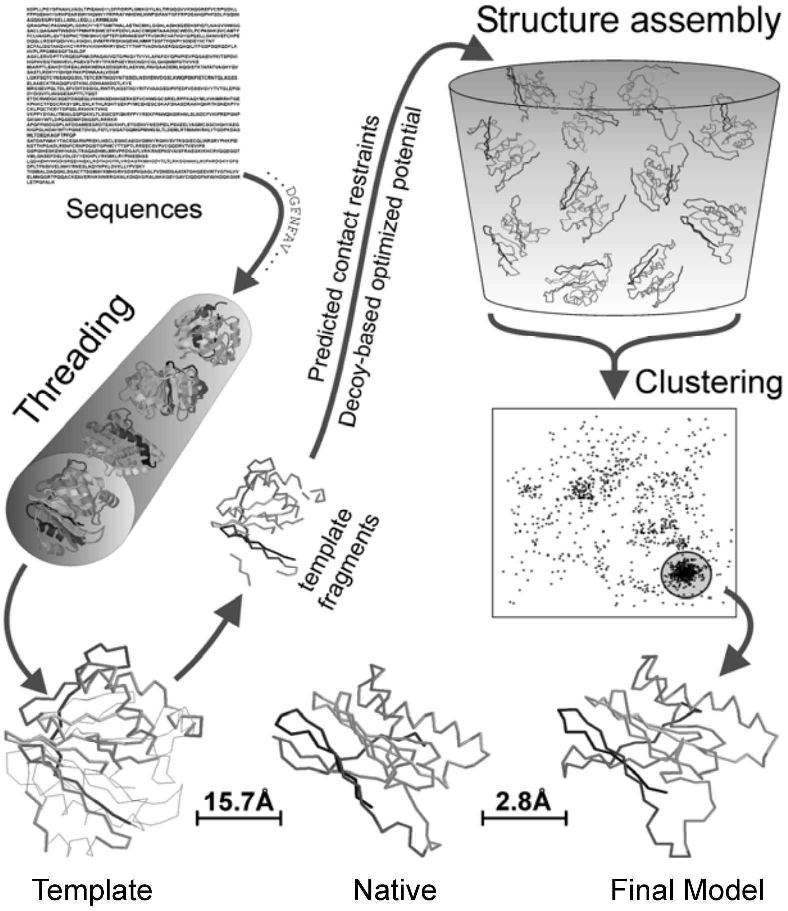
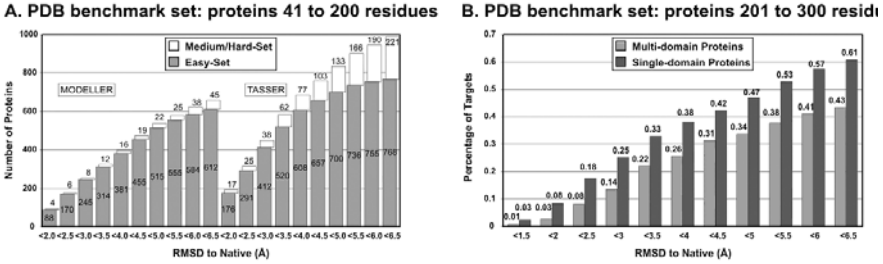

Structure assembly
Predicted contact restraints
Decoy-based optimized potential
...DGENFAV...
Sequences
Threading
template fragments
Clustering
15.7Å
2.8Å
Template
Native
Final Model

Figure 4



A. PDB benchmark set: proteins 41 to 200 residues

B. PDB benchmark set: proteins 201 to 300 resid

Figure 5

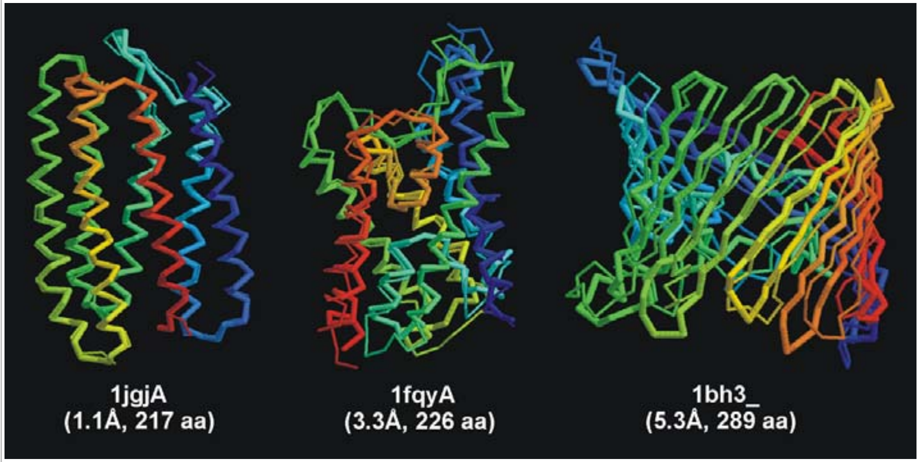
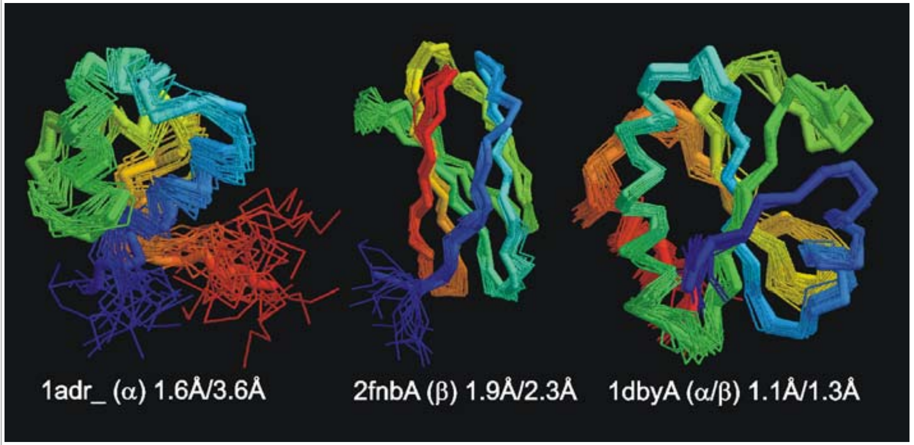


Figure 6


1adr_ (α) 1.6Å/3.6Å
2fnbA (β) 1.9Å/2.3Å
1dbyA (α/β) 1.1Å/1.3Å