Chapter 11: Genome-wide protein structure prediction

Srayanta Mukherjee[a,b], Andras Szilagyi[b,c], Ambrish Roy[a,b], Yang Zhang[a,b]*

[a]Center for Computational Medicine and Bioinformatics, University of Michigan,
100 Washtenaw Ave, Ann Arbor, MI 48109, USA
[b]Center for Bioinformatics, University of Kansas, 2030 Becker Drive, Lawrence,
KS 66047, USA
[c]Institute of Enzymology, BRC, Hungarian Academy of Sciences, Karolina út 29,
H-1113 Budapest, Hungary

Abstract

The post-genomic era has witnessed an explosion of protein sequences in the public databases; but this has not been complemented by the availability of genome-wide structure and function information, due to the technical difficulties and labor expenses incurred by existing experimental techniques. The rapid advancements in computer-based protein structure prediction methods have enabled automated and yet reliable methods for generating 3D structural models of proteins. Genome-scale structure prediction experiments have been conducted by a number of groups, starting as early as in 1997, and some noteworthy efforts have been made using the MODELLER and ROSETTA methods. Along another line, TOUCHSTONE was used to predict the structures of all 85 small proteins in the *M. genitalium* genome, which established template-refinement based structure prediction as a practical approach for genome-scale experiments. This was followed by the development of TASSER and I-TASSER algorithms which use a combination of various approaches for threading, fragment assembly, *ab initio* loop modeling, and structural refinement to predict the structures. A successful structural prediction for all medium-sized open reading frames (ORFs) in the *E. coli* genome was demonstrated by this method, achieving high-accuracy models for 920 out of 1360 proteins. GPCRs are an extremely important class of membrane proteins for which only very few structures are available in the PDB. TASSER was used to predict the structures of all 907 putative GPCRs in the human genome, and the high accuracy confirmed by newly solved GPCR structures and recent blind tests have demonstrated the usefulness and robustness of the TASSER/I-TASSER models for the functional annotation of GPCRs. Recently, the I-TASSER protein structure prediction method has been used as a basis for functional annotation of protein sequences. The increasing popularity and need for such automated structure and function prediction algorithms can be judged by the fact that the I-TASSER server has generated structure predictions for 35,000 proteins submitted by more than 8,000 users from 86 countries in the last 24 months. The success of these modeling experiments demonstrates significant new progress in high-throughput and genome-wide protein structure prediction.

*All correspondence should be addressed to E-mail: **zhng@umich.edu**

11.1 Introduction

The post-genomic era has witnessed an explosion of sequence-level information for proteins which, however, has not been complemented by the availability of structural information, mainly due to the inherent limitations of current experimental techniques for determining the protein structure. The increasing gap between the sequence and structure space (shown in Figure 1), along with the awareness that the three dimensional (3D) structure of a protein is closely linked to its biological function (Lopez *et al.* 2007), has prompted the structural genomics (SG) project to increase the throughput of experimental structure determination (Baker *et al.* 2001 ; Gerstein *et al.* 2003 ; Chandonia *et al.* 2006) and to provide a framework for inferring the biological function (Skolnick *et al.* 2000 ; Aloy *et al.* 2001) of proteins. While SG aims to structurally characterize the protein universe by an optimized combination of experimental structure determination and comparative modeling (CM), 3D structures of at least 16,000 optimally selected proteins would be required in order for CM to cover approximately 90% of protein domain families (Vitkup *et al.* 2001), and at the current rate it appears that this goal can only be achieved in ~10 years (Zhang 2009b). This underscores the need and the feasibility for genome-wide protein structure prediction by CM and other computational methods, so that 3D structural models can be built and provide insight into molecular mechanisms, thereby promoting better understanding of physiological processes and biological systems (Wiley 1998).

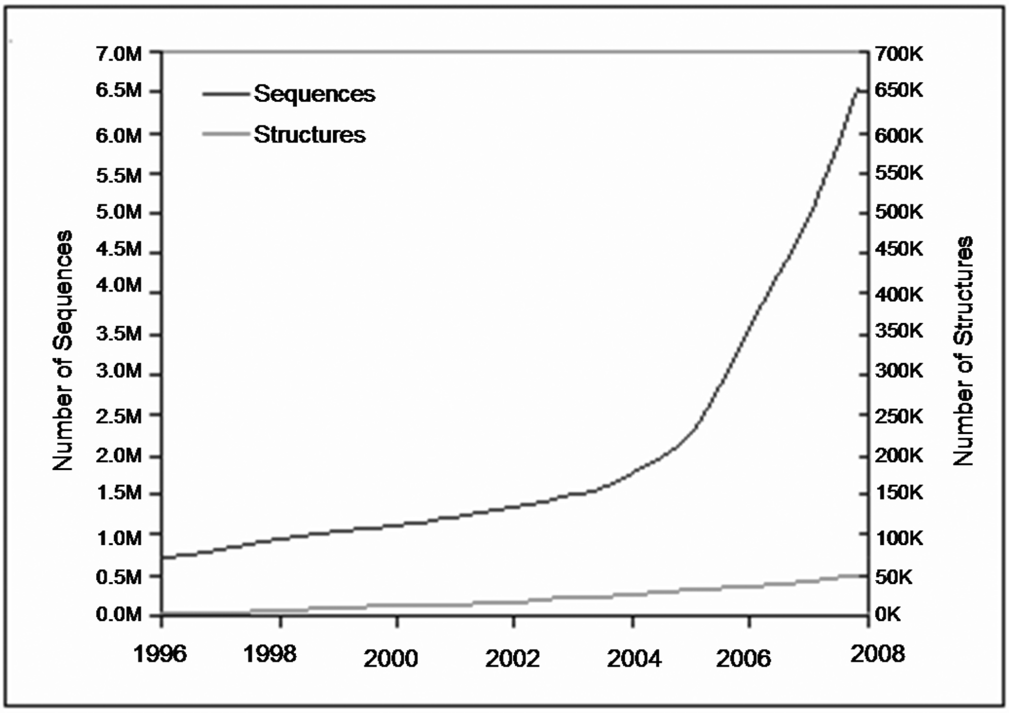

Figure 1. Plot showing the rise in the number of protein sequences in databases compared to the rise in the number of structures in the PDB (Berman *et al.* 2000) by year.

Rapid strides have been taken in the field of protein structure prediction from amino acid sequence using computational methods (Zhang 2008b). The obvious advantage of computational methods is their speed and low cost, making genome-scale structure prediction and functional annotations a reality. Protein structure prediction methods can be divided into three main categories based on the approach that is adopted (Zhang 2008b): 1) comparative or homology modeling (Sali *et al.* 1993 ; Fiser *et al.* 2000 ; Marti-Renom *et al.* 2000) 2) threading or fold recognition (Bowie *et al.* 1991 ; Jones *et al.* 1992 ; Xu *et al.* 2000 ; Skolnick *et al.* 2004b ; Wu *et al.* 2007b) and 3) *ab initio* or *de novo* methods (Kolinski *et al.* 1994 ; Simons *et al.* 1997 ; Liwo *et al.* 1999 ; Kihara *et al.* 2001 ; Zhang *et al.* 2003 ; Bradley *et al.* 2005 ; Klepeis *et al.* 2005 ; Oldziej *et al.* 2005 ; Wu *et al.* 2007a).

In comparative modeling (CM), the protein structure is constructed by matching the sequence of the protein of interest (query protein) to an evolutionarily related protein with a known structure (template protein) in the PDB. Thus, a prerequisite for comparative modeling technique is the presence of a homologous protein in the PDB (Berman *et al.* 2000) library. For proteins with >50% sequence identity to their templates, models built by CM techniques can have up to 1 Å RMSD from the native structure for the backbone atoms. For proteins which have 30 to 50% sequence identity with their template, the models often have ~85% of their core regions within an RMSD of 3.5 Å from the native structure, with errors mainly in loop regions. When the sequence identity drops below 30% (in the twilight zone (Rost 1999)), modeling accuracy sharply decreases because of substantial alignment errors and lack of significant template hits. Also, by definition, models built by CM usually have a strong bias and are closer to the template structure than the native structure of the target protein (Tramontano *et al.* 2003 ; Read *et al.* 2007).

Threading or fold recognition is similar to CM modeling in the sense that it also searches a structure library to identify a known structure which would "best fit" a given query sequence; however, an evolutionary relationship (homology) between the query and the template is not a prerequisite in this case. These "sequence to structure" alignment approaches usually employ a wide range of scoring functions to find the best alignment, and may rely on distance dependent potentials (Sippl *et al.* 1992), predicted secondary structure (McGuffin *et al.* 2003), solvent accessibility (Zhang *et al.* 1997 ; Chen *et al.* 2005a), and other predicted structural features. Most of the successful threading approaches use scores combining sequence features and predicted structural information (Skolnick *et al.* 2004b ; Zhou *et al.* 2005 ; Wu *et al.* 2008b), with a search engine of either dynamic programming (Needleman *et al.* 1970 ; Smith *et al.* 1981) or a Hidden Markov model (Karplus *et al.* 1998 ; Soding 2005) for remote homology detection and fold recognition.

Ab initio or *de novo* methods originally referred to the approaches purely based on physicochemical properties; however, some of the contemporary algorithms in this category do use evolutionary and knowledge-based information to collect spatial restraints or to detect structural fragments to assist structural assembly. However, by definition, *ab initio* methods are not dependent on the presence of known structures which are sequentially or structurally similar to a given query sequence. The guiding principle of this approach is the Anfinsen hypothesis (Anfinsen 1973), which states that the native structure of the protein lies at the global energy minimum of the configurational space. Therefore, *ab initio* approaches try to fold a given protein based on various force fields via conformational search. Though some notable developments have been made in this field (Kolinski *et al.* 1994 ; Simons *et al.* 1997 ; Liwo *et al.* 1999 ; Kihara *et al.* 2001 ; Zhang *et al.* 2003 ; Bradley *et al.* 2005 ; Klepeis *et al.* 2005 ; Oldziej *et al.* 2005 ; Wu *et al.* 2007a), predicting three-dimensional structure of proteins longer than 150 amino acids is still an unsolved problem due to the inaccuracy of available force fields and the bottlenecks arising out of insufficient conformational search.

Significant progress has been achieved in developing composite structure predictions which combine various approaches to comparative modeling, threading and *ab initio* folding. The Threading ASSEmbly Refinement (TASSER) (Skolnick *et al.* 2004a) and Iterative Threading ASSEmbly Refinement (I-TASSER) (Wu *et al.* 2007a ; Zhang 2007, 2008a) methods are notable examples in this category.

In what follows, we give an overview of the field with a focus on genome-wide automated protein structure prediction. We start with a discussion of the early attempts at large-scale structure prediction. Then, we provide an introduction to the TASSER and I-TASSER algorithms, followed by a review of the genome-scale structure prediction experiments conducted using these composite methods. Lastly, we conclude the chapter with comments on the usefulness of genome-wide structure prediction and current challenges in the field. Due to the space limit of this chapter, we are not aiming at providing an exhaustive list of efforts made in this important field.

11.2 Pioneering efforts in genome-scale structure predictions

One of the earliest attempts aiming at structure prediction on a genomic scale was carried out by Fischer and Eisenberg (1997) (Fischer *et al.* 1997) on the *Mycoplasma genitalium* genome (Fraser *et al.* 1995). The primary goal of the experiment was to assign a fold to each of the 486 putative proteins in the *M. genitalium* genome. The method used in this experiment was protein fold recognition using threading. Thus, each target sequence was threaded onto structures in a library of representative protein structures obtained from the PDB (Berman *et al.* 2000), using both sequence-profile and profile-profile alignment to find a template protein in the database of known structures that presumably has a similar structure to the target protein or at least shares a structural motif with it. Using this procedure, a fold could be assigned with high confidence to 22% of the proteins in the genome. At the time of the experiment, the threading template library included only 1,632 entries at a 50% pair-wise sequence identity cutoff.

A genome-scale structure prediction of proteins in the *Saccharomyces cerevisiae* genome was undertaken by Sanchez and Sali (1998) (Sanchez *et al.* 1998), using comparative protein structure modeling. The program MODELLER (Sali *et al.* 1993 ; Sanchez *et al.* 1997), which models an unknown protein structure based on the satisfaction of spatial restraints derived from homologous proteins of known structure, was used for all three steps necessary to perform comparative modeling, namely: sequence-structure alignment, building a model based on the restraints from templates, and evaluation of the quality of the model. Structure modeling was carried out on 6,218 ORFs, using a template library consisting of 2,045 proteins at a 95% pair-wise sequence identity cutoff. Models were generated for substantial segments of 1,071 ORFs (17.2%) from the complete genome. In contrast, only 40 proteins had been solved experimentally at that time.

Taking it one step further, Sanchez *et al.* carried out a “multi-genomic” scale comparative structure modeling for approximately 17,000 proteins from 10 complete genomes and all sequences from *Arabidopsis thaliana* and *Homo sapiens* (Sanchez *et al.* 2000) available at that time. The models were generated using the MODPIPE pipeline software (Sanchez *et al.* 1998), an integration of PSI-BLAST (Altshucl *et al.* 1997) for threading with MODELLER (Sali *et al.* 1993), and were deposited in the MODBASE database. The MODPIPE pipeline thus established a state-of-the-art automated procedure capable of performing large-scale protein structure modeling that could be used for various biological applications.

TOUCHSTONE (Kolinski *et al.* 1994 ; Kihara *et al.* 2001 ; Zhang *et al.* 2003), a Monte Carlo simulation based method built on a reduced knowledge-based force field combined with secondary structure prediction and threading-based tertiary structure restraints, was used for the genome-scale prediction of all small proteins (those shorter than 150 amino acids) in the *Mycoplasma genitalium* genome (Kihara *et al.* 2002), demonstrating the feasibility of large-scale prediction experiments using *ab initio* based modeling methods. Since the structure of none of the 85 small proteins in the genome had been solved experimentally at that time, it was not possible to compare the models with the native structures. However, as judged based on TOUCHSTONE benchmarking results on a 65-protein test set, the results were promising. Out of the 85 proteins, the threading program PROSPECTOR (Skolnick *et al.* 2001) was able to produce significant threading hits for 34 proteins, all of which were then used to produce reliable full-length models. For 29 out of the remaining 51 proteins without any significant threading hits, the Monte Carlo simulations converged to five or fewer clusters. Based on a simple application of the statistics obtained from the benchmarking study, it was concluded that the models were reliable for 24 of these 29 proteins. Thus, the total number of proteins with reliable models was estimated to be 58, or 68% of all the target proteins in the study.

Simons *et al.* conducted a large-scale test of the ROSETTA structure prediction method (Simons *et al.* 1997 ; Bradley *et al.* 2005) by predicting the structures of 150 proteins with sizes up to 150 amino acids (Simons *et al.* 2001). The protein set included 30 small (<50 residues), 127 medium-size (50 to 100 residues), and 3 relatively large proteins (>100 residues). Models with an RMSD ≤ 5Å were produced for 80% of the small proteins and 73% of the medium-size proteins. For the rest of the proteins, including the 3 large ones, the models had an RMSD between 5 to 7 Å.

In a more recent study, Malmstrom *et al.* carried out a superfamily assignment and protein structure prediction experiment on the 6,238 ORFs in the *Saccharomyces cerevisiae* genome (Malmstrom *et al.* 2007). The sequences were parsed into 14,934 structural domains, 47% of which showed detectable similarity to homologs or analogs of known structure. From the remaining 53% of the domains, the ones shorter than 150 residues and lacking predicted transmembrane helices were selected for *ab initio* structure prediction using ROSETTA. For each of the selected 3,338 domains, 10,000 models were generated by ROSETTA and then condensed to 30 representative models by clustering. This large-scale computational study was an expensive effort as it required 1,350 CPU years. The resulting structural data were integrated with existing experimental data on the function, process, and localization of the domains in order to assign them to SCOP superfamilies; an assignment was made for 581 domains. In addition, structural annotations were assigned to 7,094 domains with structures predicted using fold recognition or homology modeling. The genome-wide predictions and superfamily assignments produced by this ground-breaking study can serve as a basis for the generation of experimentally testable hypotheses about the structure-function relationships of a large number of yeast proteins.

11.3 TASSER Methods

TASSER is a composite structure prediction method developed in the Skolnick lab, (Skolnick *et al.* 2004a ; Zhang *et al.* 2004c) involving a hierarchical combination of template search by threading, followed by the assembly and rearrangement of continuous fragments excised from the templates. The protein conformation is specified in an on-and-off-lattice system with energy function integrating a number of structural restraints which are predicted from the threading templates. The on-and-off-lattice-based conformational search is used to generate thousands of conformations which are then subjected to iterative structural clustering for the selection of the final models (Zhang *et al.* 2004b).

The TASSER predictions begin by taking the amino acid sequence as input, which is then subjected to "sequence-structure alignment" or threading by PROSPECTOR_3 (Skolnick *et al.* 2004b) against a comprehensive threading library. The threading process utilizes close and distant sequence profiles and predicted secondary structure information from PSIPRED (Jones 1999) to find the best match. The alignment is performed using the Needleman-Wunsch dynamic programming algorithm (Needleman *et al.* 1970), and the raw alignment score and the alignment length are used to obtain the statistical significance (Z-score) of the alignment. The alignments on different templates are ranked by the Z-score, which is also used to classify the query protein into "easy", a "medium" or a "hard". The "hard" category basically means that no good threading template is identified in the library, and the structure will have to be largely predicted by an "*ab initio*" method.

The templates found by the threading process are divided into continuously aligned (>5 residues) and gapped regions, and placed onto the CAS (C-Alpha and Side-chain center of mass) on-and-off-lattice model. The local structure of the aligned regions remains unchanged during the simulation; their

Cα atoms are excised from the template and placed off-lattice in order to keep the fidelity of the structures. In the gapped or *ab initio* regions, Cα atoms are placed on the lattice points with a grid of 0.87 Å. The side-chain centers of mass are off-lattice for all regions. The gapped regions are first filled up using a random walk of Cα-Cα bond vectors to generate a full-length model which is subsequently subjected to the parallel hyperbolic Monte Carlo sampling (Zhang *et al.* 2002). Once again the CAS model differentiates between the on- and off-lattice atoms with regard to the movements they are subjected to. The off-lattice atoms are subjected to rigid-body translation and rotation. Care is taken to ensure that the acceptance probability of a movement is approximately the same for different fragment lengths, implemented by normalizing the amplitude of movement by the length of the fragment. On the other hand, on-lattice atoms are subjected to two- to six-bond movements and sequence shifts of multiple bonds. A pictorial representation of the CAS model is shown in Figure 2.

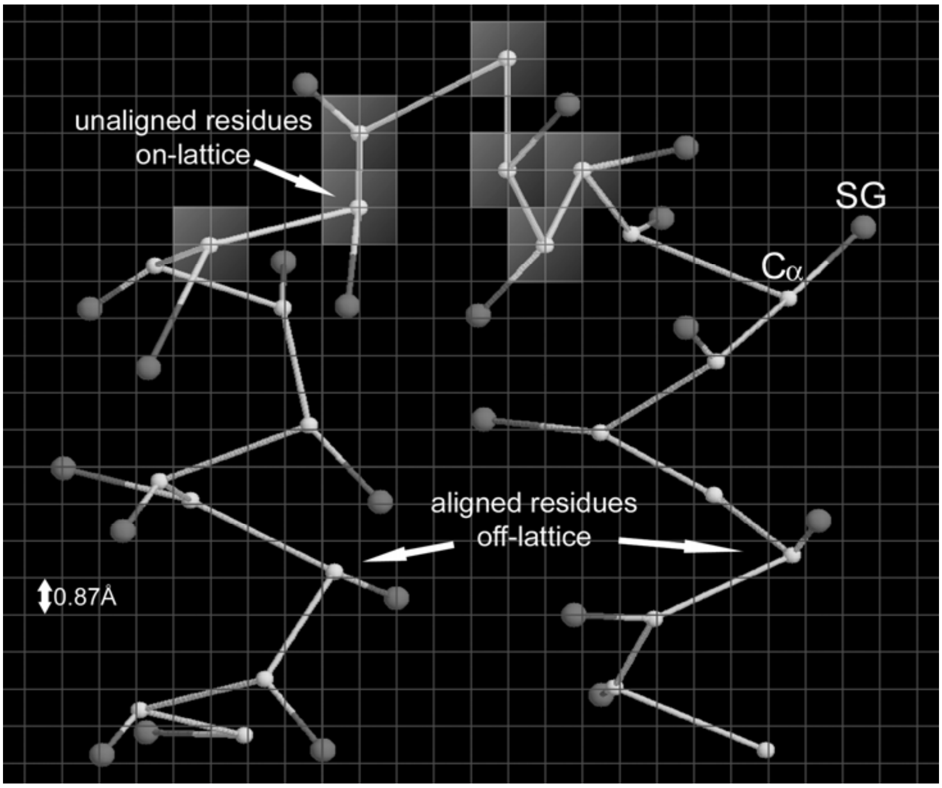

Figure 2: A schematic representation of the CAS on- and off-lattice model for a fragment of a polypeptide chain, with each residue being represented by the Cα atom and side-chain center of mass. The Cα atoms of the unaligned residues are placed on-lattice and subjected to 2-6 bond movements and multi-bond shifts. The Cα regions of the aligned regions are subjected only to rigid body rotations and translations. All side-chain atoms are off-lattice.

The TASSER energy function integrates three different classes of energy terms. The first term consists of a number of knowledge-based statistical potential derived from the PDB (Berman *et al.* 2000), including long-range side-chain pair interactions, hydrogen-bond potential terms, hydrophobic interaction and local Cα correlations. The second class includes the propensity of an amino acid to assume a particular secondary structure as predicted by PSIPRED (Jones 1999), while the third class includes protein specific tertiary structure contact restraints and a distance map calculated by PROSPECTOR_3 from the generated threading templates. The decoys generated from the TASSER sampling are finally subjected to iterative structural clustering by SPICKER (Zhang *et al.* 2004b) to rank the decoys and extract near-native final models.

11.4 I-TASSER Methods

I-TASSER is an extension of the TASSER methodology, which is implemented by running repeated iterations of the TASSER Monte Carlo sampling (Wu *et al.* 2007a). A schematic overview of the I-TASSER methodology is shown in Figure 3. The main new developments in I-TASSER compared to TASSER are: (a) LOMETS is used to extract spatial restraints from multiple threading algorithms (Wu *et al.* 2007b); (b) sequence-based contact predictions from SVMSEQ guide the *ab initio* simulations (Wu *et al.* 2008a ; Wu *et al.* 2009); (c) REMO is used to refine the hydrogen-binding network of reduced models (Li *et al.* 2009); (d) iterative TASSER reassembly (Wu *et al.* 2007a); (e) integration of structure-based functional annotations.

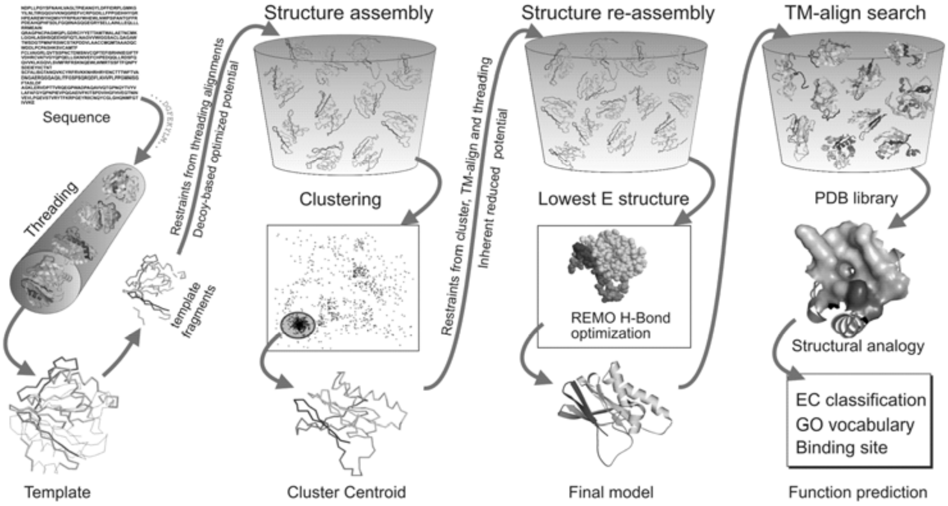

Figure 3. A schematic diagram of the I-TASSER (Wu *et al.* 2007a ; Zhang 2007, 2008a, 2009a) structure and function prediction protocol. Templates for the query protein are identified by LOMETS (Wu *et al.* 2007b), which provides template fragments and spatial restraints. The template fragments are then assembled by parallel hyperbolic Monte Carlo simulations (Zhang *et al.* 2002). The conformations generated during the simulation are clustered using SPICKER (Zhang *et al.* 2004b), in order to find the structure with the lowest free energy. As an iterative strategy, the cluster centroids are then subjected to second round of simulation with the purpose of refining the structure and removing clashes. The final all-atom models are generated by REMO (Li *et al.* 2009) through the optimization of hydrogen-bonding networks. Functional homologs (PDB structures that have an associated EC number/GO term/known binding site) of the final models are identified using both global structural search (Zhang *et al.* 2005b) and local structure alignment programs which aim at finding matches between the binding/active sites of the predicted structure and templates with known function.

The starting templates in I-TASSER are collected by LOMETS (Wu *et al.* 2007b), a meta-threading server combining 9 state-of-the-art threading algorithms: FUGUE (Shi *et al.* 2001), HHsearch (Soding 2005), MUSTER (Wu *et al.* 2008b) PROSPECT2 (Xu *et al.* 2000), PROSPECTOR3 (Skolnick *et al.* 2004b) SAM-T02 (Karplus *et al.* 1998), SPARKS2 (Zhou *et al.* 2004) SP3, (Zhou *et al.* 2005) and PPA (Wu *et al.* 2007b). On average, as tested on a benchmark set of 620 non-homologous proteins, the threading alignment found by LOMETS outperforms the best individual threading programs, with a TM-score increase of at least 8%.

The new potential terms that have been incorporated in I-TASSER include the predicted accessible surface area (ASA) (Chen *et al.* 2005a ; Wu *et al.* 2007a) and sequence-based contact predictions (Wu *et al.* 2008a). Both energy terms have been derived and optimized using machine learning methods. The overall correlation between the actual exposed area as calculated by STRIDE (Frishman *et al.* 1995) and that predicted by a neural network is 0.71, based on a test on 2,234 non-homologous proteins. In the latest version of I-TASSER (Zhang 2009a), the sequence-based pairwise residue contact information from SVMSEQ (Wu *et al.* 2008a), SVMCON (Cheng *et al.* 2007) and BETACON (Cheng *et al.* 2005) are used to constrain the simulation search and improve the funnel around the global minimum of the energy landscape.

The trajectories of the low-temperature replicas of the first-round TASSER simulations are clustered by SPICKER (Zhang *et al.* 2004b). The cluster centroids are obtained by averaging all the clustered structures after superposition and are ranked based on the structure density of the cluster. Cluster centroids generally have a number of non-physical steric clashes between Cα atoms and can be over-compressed. Starting from the selected SPICKER cluster centroids, the TASSER Monte Carlo simulation is performed again (see Figure 3). While the inherent I-TASSER potential remains unchanged in the second run, external constraints are added, which are derived by pooling the initial high-confidence restraints from threading alignments, the distance and contact restraints from the combination of the centroid structures, and the PDB structures identified by the structure alignment program TM-align (Zhang *et al.* 2005b) using the cluster centroids as query structures. The conformation with the lowest energy in the second round is selected as the final model. The main purpose of this iterative strategy is to remove the steric clashes of the cluster centroids. To increase the biological usefulness of protein models, all-atom models are generated by REMO (Li *et al.* 2009) simulations, which include three general steps: (1) removal of steric clashes by moving around each of the C_α atoms that clash with other residues; (2) backbone reconstruction by scanning a backbone isomer library collected from the solved high-resolution structures in the PDB library; (3) hydrogen bonding network optimization based on predicted secondary structure from PSIPRED. Finally, Scwrl3.0 (Canutescu *et al.* 2003) is used to add the side-chain rotamers.

Recently, I-TASSER was extended by an additional component to predict the biological function of the query proteins. The procedure involves matching the I-TASSER-generated structural models against representative libraries of proteins with known function using both global and local structure alignment based methods in order to find the best functional homologs in the PDB library. Based on a large-scale benchmark test set of 218 non-homologous proteins, it was found that even when the structures are predicted after removing all the homologous templates from the template library, the correct function (EC number and GO terms) could be predicted for 72% of the test proteins with a precision of 74% (Roy *et al.* 2010).

11.5 TASSER/I-TASSER structure prediction on large-scale benchmarks

For a comprehensive test of the methodology, we collected a representative set of 2,234 single-domain proteins in the PDB whose size ranged from 41 to 300 residues (Skolnick *et al.* 2004a ; Zhang *et al.* 2004c). For each protein, homologous templates with sequence identity >30% to the target are excluded from the threading template library (Skolnick *et al.* 2004a ; Zhang *et al.* 2004c). In Figure 4a, we present a chart showing the fractions of I-TASSER-generated models having RMSDs (from the native structure) below various thresholds. About 2/3 of targets (1,470/2,234) have an acceptable topology (RMSD from native <6.5 Å); 46% of targets (1,026/2,234) have an RMSD from the native structure <4 Å. As the RMSD threshold decreases, the fraction of models below the threshold (especially those <2Å) sharply drops, which is partially due to the limitations of the TASSER potential with regard to high-resolution modeling.

Because there is no template alignment available, loop modeling is a difficult unsolved problem in protein structure prediction. In the 2,234-protein benchmark set, there are overall 3,565 unaligned regions (ranging from 4 to 170 residues long, mainly in loops and tails). If we assess loop modeling accuracy by calculating the RMSD between the predicted and the native loop conformations based on a superposition of the stem residues (Zhang *et al.* 2004c), the average RMSD for all 3,565 loops is 6.1 Å, a low value in comparison with the RMSD of 14.2 Å obtained when the same loops are built by MODELLER (Sali *et al.* 1993). If we use an RMSD cutoff of 4 Å to define success, MODELLER succeeds in 14% (499 of 3,565) of the cases, whereas TASSER *ab initio* modeling is successful in 44% (1,567 of 3,565) of the cases.

In Figure 4b, we show the RMSD distribution of I-TASSER *ab initio* models for 56 small (<120 residues) single-domain proteins (Wu *et al.* 2007a). Any meaningful template with a sequence identity >20% to the target or having a PSI-BLAST E-value <0.5 was excluded. In this limit, about 90% of the final models have a correct fold, with an RMSD <6.5 Å from the native structure. The average RMSD of the I-TASSER models is 3.8 Å, compared to 5.9 Å by TOUCHSTONE (Zhang *et al.* 2003) for the same set of proteins. Since the template exclusion used here is much stricter than that in TOUCHSTONE, which used a sequence identity cutoff of 30%, these data demonstrate significant progress by I-TASSER over TOUCHSTONE in *ab initio* modeling. For the 16 proteins which were also tested by Bradley et al. (Bradley *et al.* 2005), the overall result is comparable with that from all-atomic ROSETTA simulations (both have an average RMSD of 3.8 Å), but the CPU time required by I-TASSER was much shorter (5 CPU hours vs. 150 CPU days).

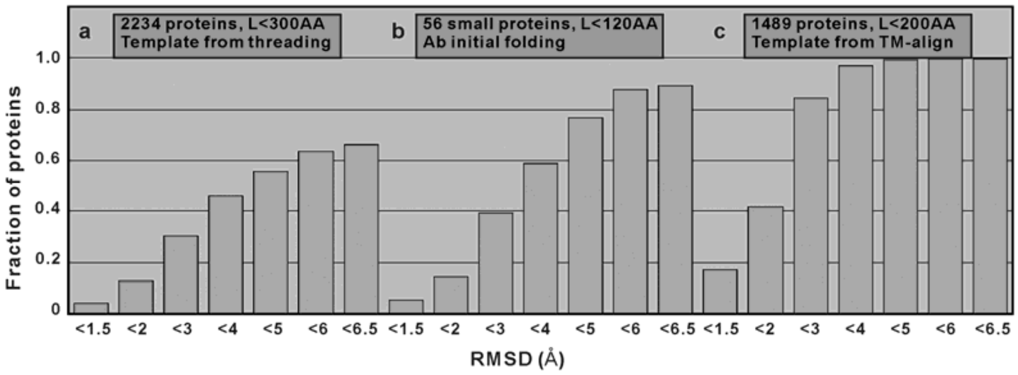

Figure 4. The success rate of TASSER/I-TASSER on three benchmark sets vs. the RMSD threshold defining success. **(a)** 2,234 proteins with non-homologous templates from the threading program (Skolnick *et al.* 2004a ; Zhang *et al.* 2004c). **(b)** 56 small proteins in the *ab initio* limit (Wu *et al.* 2007a). **(c)** 1,489 proteins with non-homologous templates from structure alignment by TM-align (Zhang *et al.* 2005b ; Zhang *et al.* 2005a)

In Figure 4c, we show the success rates of a procedure that uses the best templates identified by structural alignments. First, a representative set of 1,489 target proteins from the PDB with lengths between 41 and 200 residues was taken as target proteins; then the native structure of each target was superimposed to structures in the PDB to identify the best template, while homologous templates with sequence identity >25% to the target are excluded (Zhang *et al.* 2005a). The purpose of this experiment was to examine whether the current PDB structure library is complete (Zhang *et al.* 2005a) and if so, how well TASSER structure prediction can perform when starting from the best possible non-homologous templates. The data show that starting from structural alignments, TASSER generates "foldable" models with a RMSD <6 Å for almost all targets, and "good" models with a RMSD <4 Å for 97% of the targets. Although the fraction of high-resolution models (<2 Å) is still relatively low, these striking data suggest that when the goal is to build a model with correct topology (RMSD <6 Å (Zhang *et al.* 2005a)), the structure prediction problem for single-domain proteins could in principle be solved using the current PDB by efficient fold recognition algorithms that would be able to recover the structural alignments (Zhang *et al.* 2005a). Indeed, all single-domain folds in the PDB are represented in an artificially generated library of compact, hydrogen-bonded, sticky homo-polypeptide structures (Zhang *et al.* 2006b).

11.6 Prediction of all medium-sized ORFs in the *E.coli* genome

Inspired by the success of the benchmark test, a genome-scale structure prediction experiment (Zhang *et al.* 2004a) was carried out for all 1,360 medium-sized ORFs (<201 residues) in the *E. coli* genome (Blattner *et al.* 1997). The PROSPECTOR_3 threading algorithm assigns 829 proteins to the easy set, 521 to the medium set and only 10 to the hard set; this target distribution is quite similar to that of the benchmark set. Based on the benchmarking study described above, a confidence score (or C-score) was defined to assess the quality of a model, which is a combination of the Z-score of the threading template and the degree of convergence of the conformations generated by the CAS refinement simulations. The confidence score is defined by

$$C\text{-}score = \ln\left(\frac{M}{\langle rmsd\rangle M_{tot}} Z\right) \qquad (1)$$

where M is the multiplicity of structures in a given SPICKER cluster, <*rmsd*> is the average RMSD of the structures in the cluster from the cluster centroid, M_{tot} is the total number of conformations used as input to SPICKER, and Z is the Z-score of the starting template. Having observed a good correlation of C-score with RMSD for the benchmark set (a C-score > -1.5 is roughly equivalent to a TM-score > 0.5 which indicates a similar fold), this C-score could be used to assess the quality of the models generated for the *E. coli* ORFs. The *E. coli* ORFs were found to have a C-score distribution similar to the one observed for the PDB benchmark set. If the correlation between C-score and RMSD is assumed to be the same for the *E. coli* set as the benchmark set, ~920 or 68% of the models generated can be considered to be reliable. The percentage of correct structures is slightly higher than for the PDB benchmark set (see Figure 4), partly because homologous proteins were not excluded during the threading process for the *E. coli* ORFs. A histogram showing the distribution of C-scores for the *E. coli* ORFs and the PDB benchmark set is shown in Figure 5.

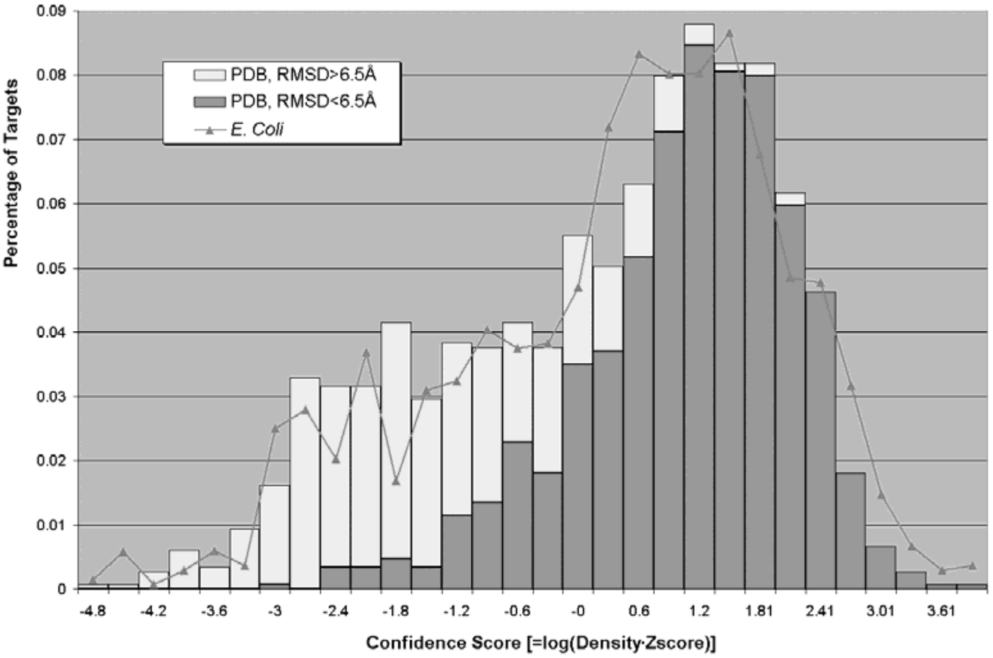

Figure 5: A histogram showing the C-score (defined in Eq. 1) distribution of models for the *E coli* genome (solid line) and the PDB benchmark set (bars). The different colors in the bars indicate the fraction of targets below and above an RMSD cutoff of 6.5Å for the PDB benchmark set.

Transmembrane proteins are particularly difficult to crystallize, with difficulties ranging from expression of membrane proteins in microbial host cells to purification of the protein to the crystallization process itself, due to the amphipathic nature of their environment (Ostermeier *et al.* 1997 ; Caffrey 2003). Hence, accurate prediction of membrane proteins is of special importance. According to MEMSAT (Jones *et al.* 1994), 309 of the 1,360 *E. coli* ORFs belong to the membrane protein class. The TASSER models generated for these ORFs share good consistency with the MEMSAT predictions in having at least one long, putative transmembrane helix. If the C-score, defined above, is used to map the models, 174 of the 309 proteins or 56% have a probability >60% to have an overall RMSD <6.5Å and about 146 or 47% have a chance >80% to have an RMSD less than 6.5Å.

11.7 Structural modeling of all 907 putative GPCRs in the human genome

G protein-coupled receptors (GPCRs) comprise the largest family of integral membrane proteins and act as cell surface receptors responsible for the transduction of an endogenous signal into a cellular response (Watson *et al.* 1994 ; Flower 1999). Many diseases involve their malfunction, making them the most important class of drug targets (Flower 1999 ; Drews 2000 ; Lundstrom 2005 ; Hubbard 2006) However, structure-based drug design has been hampered by the lack of atomic-level protein structure information for GPCRs. Until now, only four GPCR structures, bovine rhodopsin, (Palczewski *et al.* 2000) turkey β_1-adrenergic receptor, (Warne *et al.* 2008) and human β_2-adrenergic (Cherezov *et al.* 2007 ; Rasmussen *et al.* 2007 ; Rosenbaum *et al.* 2007) and A_{2A}-adenosine receptors, (Jaakola *et al.* 2008) have been solved.

We collected 907 human GPCRs from the registered entries at http://www.cmbi.kun.nl/7tm/htmls/entries.html and http://www.expasy.org/cgi-bin/lists?7tmrlist.txt. TASSER was employed to model all the GPCRs (Zhang *et al.* 2006a). The resulting models are publicly downloadable from http://cssb.biology.gatech.edu/ skolnick/files/gpcr/gpcr.html.

Because there was no restraint on the global topology, it is of interest to examine how often the models adopt a typical TM-helix bundle architecture. Using an automatic TM-helix identification program, we found that 862 of the 907 GPCRs have the 7-helix bundle topology, although only 744 targets started from a TM-helix-like template. Among the other 45 cases, 16 are incomplete or alternatively spliced transcripts; most are missing the majority of their TM regions; three (Q8TDU0, Q8TDV3, Q96HT6) do not appear to be GPCRs based on sequence analysis; (Marchler-Bauer *et al.* 2005) two (Q9HC23 and P06850) are wrongly annotated as GPCRs (Pisarska *et al.* 2001 ; Chen *et al.* 2005b). The remainder may represent incorrect TASSER predictions, since the C-score of these targets is low.

Although at the time of the study, there was no solved X-ray or NMR structure for any human GPCR and a direct comparison of models with experimental structures was not possible, two criteria were found to be useful for the assessment our models. First, we use the model's C-score (see Eq. 1). Based on the 2,234-protein benchmark set, the correlation coefficient between C-score and RMSD is 0.85 (Skolnick *et al.* 2004a), with a similar correlation also obtained for a benchmark set of 38 membrane proteins (Zhang *et al.* 2006a). Due to the uniform 7-TM topology and the robust sequence profiles (Skolnick *et al.* 2004b), a much higher fraction of the GPCR models have a high C-score than the models generated for PDB benchmark set (Figure 6). Assuming that the GPCR models have the same correlation between C-score and RMSD as those of the PDB benchmark proteins, we estimate that 819 GPCR models have a correct fold with a RMSD below 6.5 Å.

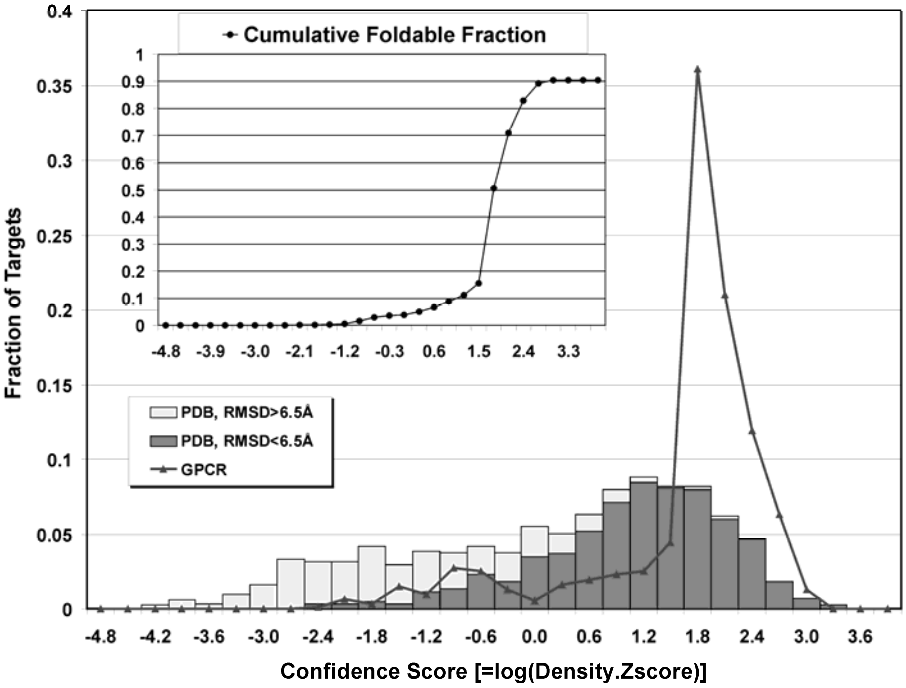

Figure 6: Histogram showing the distribution of C-scores (defined in Eq. 1) for the PDB benchmark set (bars) and the GPCR models (solid line). The different colors in the bars indicate the fraction of models in the PDB benchmark set with an RMSD below (dark grey) and above (light gray) 6.5Å, respectively.

Second, we evaluated the GPCR models by considering the affinity labeling and site-directed mutagenesis experiments designed to identify critical residues and motifs that participate in ligand binding (Schwartz 1994 ; Flower 1999 ; Shi *et al.* 2002). These data provide useful clues about the spatial arrangements of binding site residues, and we can examine if our models are consistent with these. We checked all TASSER models with C-score >1.3 with available site-directed mutagenesis data collected from 64 papers. These included angiotensin receptor 1, chemokine receptors, opioid receptors, thromboxane A2 receptor, neuromedin B receptor, melatonin 2 receptor, gonadotropin-releasing hormone receptor, and neuropeptide Y receptors. Excluding N- and C-terminal tails, the TASSER-predicted models were consistent with the experiment (Zhang *et al.* 2006a). Figure 7 shows the human Y1 receptor. Consistent with the mutagenesis studies (Zhou *et al.* 1994 ; Hwa *et al.* 1995 ; Sautel *et al.* 1996 ; Du *et al.* 1997), the ligand binding residues are well grouped in the model.

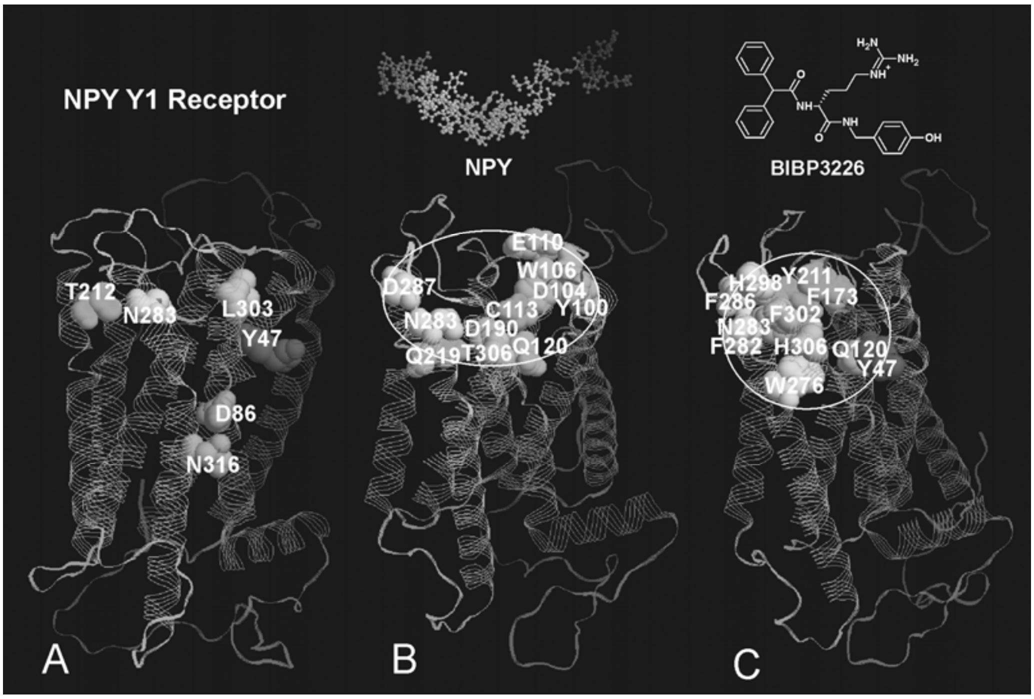

Figure 7. The first model of neuropeptide Y Y1 receptor predicted by TASSER, having a C-score of 1.93. (Zhang *et al.* 2006a) Secondary structure elements are displayed as open ribbons. (**A**). Three pairs of highlighted residues are in contact as verified by the reciprocal mutagenesis experiments. (**B**). Highlighted residues represent the important residues identified in NPY agonist binding mutagenesis experiments. (**C**). Highlighted residues are the critical residues identified by BIBP3226 antagonist binding mutagenesis experiments. (Sautel *et al.* 1996 ; Du *et al.* 1997)

Based on an all-against-all comparison of the predicted structures, GPCRs in the same functional family were found to be more conserved in structure space than in sequence space. This finding establishes the possibility of functional annotation of orphan proteins based on topology-level comparisons of predicted structures. One such instance is the RDC1 receptor, which was considered an orphan receptor for 15 years; its closest but weak relative is the adrenomedullin receptor (AMDR) based on phylogenetic studies (Ladoux *et al.* 2000). The TASSER structural predictions placed the RDC1 receptor in the family of chemokine receptors because the predicted RDC1 structure is closest to the predicted structure of the CXCR4 chemokine receptor (Zhang *et al.* 2006a). This finding was later confirmed by binding experiments. (Miao *et al.* 2007)

After the modeling had been done (Zhang *et al.* 2006a), the structures of two human GPCRs, the β_2 adrenergic and A2A adenosine receptor, were solved by two laboratories at Stanford University and The Scripps Research Institute (Cherezov *et al.* 2007 ; Rasmussen *et al.* 2007 ; Jaakola *et al.* 2008). These structures provide a unique opportunity to objectively examine the quality of the TASSER models. β$_2$AR is a class-A receptor that is 413 residues long. It is found in human smooth muscle and mediates the catecholamine-induced activation of adenylate cyclase through the action of G proteins. Efforts to crystallize wild-type β$_2$AR had been unsuccessful because of the inherent conformational plasticity mainly induced by the C-terminal tail and the third unstructured intracellular loop (ICL3) (Granier *et al.* 2007 ; Rosenbaum *et al.* 2007). To increase crystal contacts, Rasmussen et al (Rasmussen *et al.* 2007) remove the C-terminus and bind a monoclonal antibody (Mab5) to ICL3. Using high-brilliance microcrystallography, the structure of a 216 residue portion was determined at a resolution of 3.4 Å (PDB ID: 2r4rA). Cherezov et al (Cherezov *et al.* 2007) replaced ICL3 with T4 lysozyme (T4L) to increase the TM conformational stability. This led to a high-resolution structure of 282 residues with a resolution of 2.4 Å (PDB ID: 2rh1A). The missing parts are mainly from the N- and C-termini and the ICL3 region. This is the first solved human GPCR structure. Because it is longer and has a higher resolution than 2r4rA, we compared our models to 2rh1A in our analysis.

In our modeling of β$_2$AR, PROSPECTOR3 identified bovine rhodopsin (1f88A) as the template with a high significance score (Z-score=23.1). The RMSD of the 253 aligned residues from the template to the native structure is 4.94 Å with a TM-score=0.71. In the 7 TM-helix regions, i.e. TM1 (29-60), TM2 (67-96), TM3 (103-136), TM4 (147-171), TM5 (197-229), TM6 (267-298), and TM7 (305-328), the RMSD for the rhodopsin template is 3.7 Å. TASSER takes the restraints from the template and reassembles the fragments with loops built by *ab initio* modeling. As a result, the structure of the first model has an RMSD of 4.37 Å in the threading aligned regions; for the 7 TM-helix regions, the RMSD of the first TASSER model is 2.28 Å (Figure 8, left panel). For the full-length model, the RMSD to native is 4.88 Å with a TM-score=0.82.

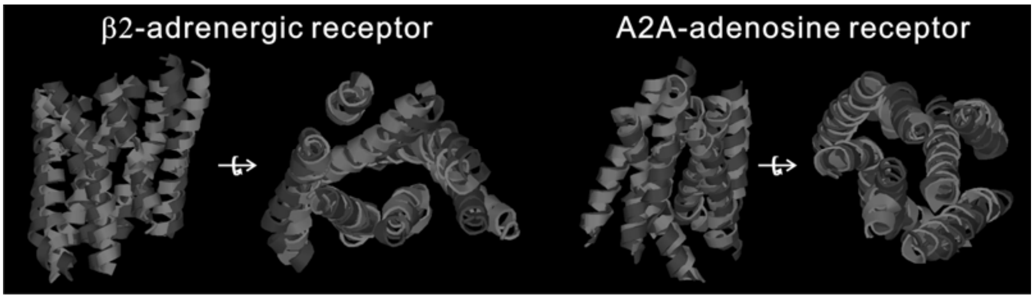

Figure 8. Side and top views of the first TASSER model (gray) superposed on the crystal structure (dark) of β$_2$AR and ADORA2A over the seven transmembrane regions with a RMSD 2.28 Å and 2.87 Å, respectively.

ADORA2A is a class-A purinergic receptor with a length of 412 residues. Stevens and coworkers exploited a similar T4L fusion strategy to crystallize the receptor resulting in a structure of 2.6 Å resolution (PDB ID: 3eml) (Jaakola *et al.* 2008). PROSPECTOR3 identified again the bovine rhodopsin as a template with an RMSD of 5.13 Å in 262 aligned residues; the RMSD of the templates in the TM-regions is 3.23 Å. After TASSER reassembly, the RMSD of the first model in the threading aligned regions is 4.20 Å while in the 7 TM-helix regions, the RMSD of the model is reduced to 2.84 Å (Figure 8, Right panel). The overall RMSD of the full-chain model is 4.76 Å with a TM-score=0.80. It should be mentioned that the modeling here was made using rhodopsin as template. When using the newly solved adrenergic receptors which are structurally closer to ADORA2A, the models could be further improved, e.g. the RMSD in TM-helices of our model by I-TASSER which was recently submitted to the community-wide GPCR docking experiment (Michino *et al.* 2009) was 2.04 Å (model ID: 1800_2.pdb, picture not shown).

Both blind-test examples (β_2AR and ADORA2A) show that the TASSER/I-TASSER fragment assembly procedure can draw the template significantly closer to the native structure (i.e. by 1.4Å/0.6Å and 0.4Å/0.9Å in TM-helix/aligned regions for β_2AR and ADORA2A, respectively). This ability to refine structures is particularly important for modeling those GPCRs that do not have close templates in the PDB.

Currently, efforts are under way to extend the I-TASSER methodology for predicting the structure of all classes of integral membrane proteins. In an initial benchmarking study, 88 integral membrane proteins (66 α- and 24 β-barrel proteins) belonging to 24 different families were selected from the PDB and modeled using the current I-TASSER protocol, excluding any templates having >30% sequence identity with the target. Overall, 61 proteins were classified as easy targets, 24 as medium and 3 as hard targets, based on the LOMETS threading alignment. For 37 proteins, the best identified template was itself a membrane protein, and 43 templates had a TM-score >0.6 with the native target structure, showing that good templates exist in the current PDB library for ~45% of the membrane proteins. After generating full-legnth models by the I-TASSER procedure, 37 proteins in the benchmark set were modeled with an average RMSD of 4.203 Å, and 43 proteins had an average TM-score of 0.7726 for the full length model. Much effort is being made to develop a membrane-protein specific version of I-TASSER, which would be capable of taking into consideration the uniqueness of the membrane environment and predicting integral membrane protein structures even when no good template is identified in the template library, with an equivalent precision and accuracy to that for globular proteins.

11.8 Application of I-TASSER to the *Chlamydia trachomatis* genome

Bacteria from the *Chlamydia* genus are implicated in a large number of human diseases, including glaucoma and ectopic pregnancy among many others. The lack of a gene transfer system for these bacteria makes them difficult to study *ex vivo* and has greatly hampered our understanding of their biology. Although the genome sequence of many Chlamydia species are freely available in genome databases, the functional annotations of ORFs in these genomes, based on sequence comparisons has been limited due to the lack of reliable sequence similarity with proteins of known function. As residues located far apart in the sequence may be very close in 3D space, and only a few spatially conserved residues are generally responsible for a protein's function (Wallace *et al.* 1996 ;

Kleywegt 1999), predicted 3D structures for proteins from such organisms can provide meaningful insights into the key component(s) of their functionality.

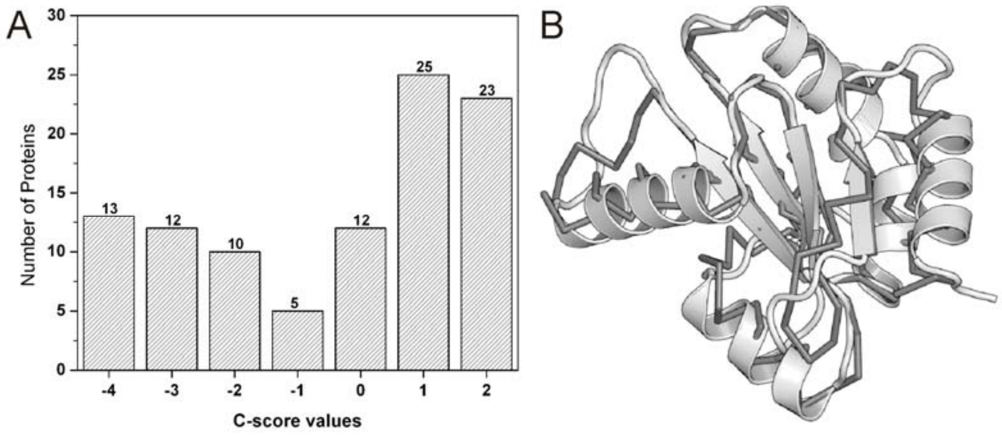

Figure 9. (A) Distribution of C-scores of predicted structures for 100 proteins in *Chlamydia trachomatis* genome using I-TASSER. (B) Comparison of the modeled structure of CT780 (dark grey and stick) with the crystal structure of thioredoxin disulphide isomerase (dark grey and cartoon) from *Chlamydia pneumoniae*.

The I-TASSER methodology for protein structure and function prediction was recently applied to 100 ORFs with no functional annotation in the *Chlamydia trachomatis* genome. Figure 9A shows the distribution of the confidence score (C-score) of the first I-TASSER models for all 100 proteins. Based on the correlation data of C-score with RMSD and TM-score (Zhang 2008a), it can be expected that 66 of these 100 proteins could be correctly folded (predicted TM-score >0.6) and could provide meaningful insight into the function of these proteins.

Moreover, by using a local and global structure alignment based method, a highly confident function prediction (based on a benchmark test of 218 proteins) could be made for 12 enzymatic and 38 non-enzymatic proteins, i.e. altogether exactly 50% of all target proteins. Figure 9B shows an illustrative example, the protein CT780. The structure of an ortholog of this protein, in *C. pneumoniae,* had already been solved (PDB: 2ju5). For testing purposes, the structure of CT780 was modeled by excluding this template and all other proteins having a sequence >40% with the target. The first model generated by I-TASSER had a TM-score of 0.84 in the core region (when compared to 2ju5, the *C. pneumoniae* ortholog), reflecting that the structure was predicted correctly. Based on this predicted model, TM-align identified a correct functional homolog, the third thioredoxin domain of protein disulfide isomerase A4 from mouse, with EC: 5.3.4.1 (2dj3A). Primary sequence comparison supports the annotation of the protein as a thioredoxin disulfide isomerase DsbH. Functional studies on DsbH demonstrated that it exhibits many of the enzymatic properties of thioredoxin from *E. coli* (Mac *et al.* 2008).This identified homolog shares a sequence identity of 27.1% with the query protein, showing that even when only remotely homologous templates are available, the modeled structure can provide meaningful insight into the molecular function, and can make genomic-scale functional annotation a reality.

11.9 Concluding remarks

Genome-wide structure predictions have been carried out by state-of-the-art methods for a number of organisms, with representative examples including the predictions for *Saccharomyces cerevisiae* by MODELLER (Sanchez *et al.* 1998), the yeast proteome by ROSETTA (Malmstrom *et al.* 2007), and the *E. coli* proteome (Skolnick *et al.* 2004a), all human GPCRs (Zhang *et al.* 2006a), and the *Chlamydia trachomatis* proteome by TASSER/I-TASSER (Skolnick *et al.* 2004a ; Wu *et al.* 2007a). A large percentage of the proteins in proteomes (e.g. 47% of yeast proteins) can be classified as a comparative modeling or fold recognition target, for which reliable structures can be built by current template-based methods. These predictions are immediately useful for function prediction and for the design and interpretation of wet-lab experiments (Zhang 2009b). For the proteins with no recognizable relationship to known structures, *ab initio* methods have to be developed for structure prediction. However, there are serious limitations to the application of the *ab initio* methods, which hamper their use in genome-wide prediction. In the absence of recognizable similarity to proteins with known domains, splitting long sequences into domains can be done with only limited accuracy. Even when domain parsing is successful, *ab initio* methods can hardly be applied to domains >150 residues (Zhang 2008b). Membrane proteins are another group which is often excluded from the prediction attempts except for special classes where homologous or analogous templates are available. Therefore, the efforts of genome-wide prediction based on *ab initio* approaches (Kihara *et al.* 2002 ; Skolnick *et al.* 2004a ; Malmstrom *et al.* 2007) and those aimed at membrane proteins (Zhang *et al.* 2006a) are exceptionally important. Although their results so far are encouraging, when all structure prediction approaches are combined, a significant fraction (~1/3) of the proteome remains that is inaccessible to current methods (Zhang 2008b).

Although the ultimate purpose of structure prediction is to help design and interpret experiments, the accuracy of the final model determines its possible use. Only high-resolution models can be used for reliable docking or drug design; the lower-resolution structures can be useful for superfamily assignment or putative functional annotation (Zhang 2009b). The refinement of low-resolution models to achieve higher resolution is therefore of key importance but remains a challenge (Kopp *et al.* 2007 ; Read *et al.* 2007).

In the context of genome-wide protein structure prediction, the “sequence-to-structure-to-function” paradigm does not necessarily have to be conceived as a one-way path. Obviously, functional annotation can be based on predicted structures; but this relationship also works the other way: existing functional information can help select the most likely structure when several different candidate structures are available. The study of Malmstrom et al. (Malmstrom *et al.* 2007) represents a prime example for this logic: the assignment of SCOP superfamilies to *ab initio* predicted domain structures was augmented by the available functional data. In the future, the integration of computational and experimental findings will be essential to enhance our understanding of biological processes.

Acknowledgements

The project is supported in part by the Alfred P. Sloan Foundation, NSF Career Award (DBI 0746198), and the National Institute of General Medical Sciences (R01GM083107, R01GM084222).

References

Aloy P, Querol E, Aviles F, Sternberg J. 2001. 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. Journal of Molecular Biology 311(2):395-408.

Altschucl S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. 1997. Gapped BLAST and PSI_BLAST: a new generation of protein database search programs. Nucleic Acids Research 25(17):3389-3402.

Anfinsen CB. 1973. Principles that govern the folding of protein chains. Science 181(96):223-230.

Baker D, Sali A. 2001. Protein structure prediction and structural genomics. Science 294:93-96.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res 28(1):235-242.

Blattner F, III GP, Bloch C, Perna N, Burland V, Riley M, Collado-Vides J, Glasner J, Rode C, Mayhew G and others. 1997. The complete genome sequence of E. coli K-12. Science 277:1453-1474.

Bowie JU, Luthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253(5016):164-170.

Bradley P, Misuara K, Baker D. 2005. Towards high-resolution de novo structure prediction for small proteins. Science 309:1868-1871.

Caffrey M. 2003. Membrane protein crystallization. J Struct Biol 142(1):108-132.

Canutescu AA, Shelenkov AA, Dunbrack RL, Jr. 2003. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 12(9):2001-2014.

Chandonia J, Brenner S. 2006. The Impact of structural genomics: expectations and outcomes. Science 311:347-351.

Chen H, Zhou HX. 2005a. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res 33(10):3193-3199.

Chen J, Kuei C, Sutton S, Wilson S, Yu J, Kamme F, Mazur C, Lovenberg T, Liu C. 2005b. Identification and pharmacological characterization of prokineticin 2beta as a selective ligand for prokineticin receptor 1. Mol Pharmacol 67(6):2070-2076.

Cheng J, Baldi P. 2005. Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms. Bioinformatics 21 Suppl 1:i75-84.

Cheng J, Baldi P. 2007. Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics 8:113.

Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK and others. 2007. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science 318(5854):1258-1265.

Drews J. 2000. Drug discovery: a historical perspective. Science 287(5460):1960-1964.

Du P, Salon JA, Tamm JA, Hou C, Cui W, Walker MW, Adham N, Dhanoa DS, Islam I, Vaysse PJ and others. 1997. Modeling the G-protein-coupled neuropeptide Y Y1 receptor agonist and antagonist binding sites. Protein Eng 10(2):109-117.

Fischer D, Eisenberg D. 1997. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. Proceedings of The National Academy of Science 94:11929-11934.

Fiser A, Do RK, Sali A. 2000. Modeling of loops in protein structures. Protein Sci 9(9):1753-1773.

Flower DR. 1999. Modelling G-protein-coupled receptors for drug design. Biochim Biophys Acta 1422(3):207-234.

Fraser C, Gocayne J, White O, Adams M, Clayton R, Fleischmann R, Bult C, Kerlavage A, Sutton G, Kelley J and others. 1995. The minimal gene complement of Mycoplasma genitalium. Science 270:397-403.

Frishman D, Argos P. 1995. Knowledge-based protein secondary structure assignment. Proteins 23(4):566-579.

Gerstein M, Edwards A, Arrowsmith C, Montelione G. 2003. Structural genomics: Current progress. Science 299(5613):1663.

Granier S, Kim S, Shafer AM, Ratnala VR, Fung JJ, Zare RN, Kobilka B. 2007. Structure and conformational changes in the C-terminal domain of the beta2-adrenoceptor: insights from fluorescence resonance energy transfer studies. J Biol Chem 282(18):13895-13905.

Hubbard R, editor. 2006. First ed: Royal Society of Chemistry.

Hwa J, Graham RM, Perez DM. 1995. Identification of critical determinants of alpha 1-adrenergic receptor subtype selective agonist binding. J Biol Chem 270(39):23189-23195.

Jaakola VP, Griffith MT, Hanson MA, Cherezov V, Chien EY, Lane JR, Ijzerman AP, Stevens RC. 2008. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. Science 322(5905):1211-1217.

Jones D. 1999. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 292:195-202.

Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. Nature 358(6381):86-89.

Jones DT, Taylor WR, Thornton JM. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 33(10):3038-3049.

Karplus K, Barrett C, Hughey R. 1998. Hidden markov models for detecting remote protein homologies. Bioinformatics 14(10):846-856.

Kihara D, Lu H, Kolinski A, Skolnick J. 2001. TOUCHSTONE: An ab initio protein structure prediction method that uses threading based tertiary restraints Proceedings of The National Academy of Science 98:10125-10130.

Kihara D, Zhang Y, Lu H, Kolinski A, J. S. 2002. Ab initio protein structure prediction on a genomic scale: application to *Mycoplasma genitalim* genome. Proceedings of The National Academy of Science 99:5993-5998.

Klepeis JL, Wei Y, Hecht MH, Floudas CA. 2005. Ab initio prediction of the three-dimensional structure of a de novo designed protein: a double-blind case study. Proteins 58(3):560-570.

Kleywegt GJ. 1999. Recognition of spatial motifs in protein structures. J Mol Biol 285(4):1887-1897.

Kolinski A, Skolnick J. 1994. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 18(4):338-352.

Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. 2007. Assessment of CASP7 predictions for template-based modeling targets. Proteins 69 Suppl 8:38-56.

Ladoux A, Frelin C. 2000. Coordinated Up-regulation by hypoxia of adrenomedullin and one of its putative receptors (RDC-1) in cells of the rat blood-brain barrier. J Biol Chem 275(51):39914-39919.

Li Y, Zhang Y. 2009. REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. Proteins.

Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA. 1999. Protein structure prediction by global optimization of a potential energy function. Proc Natl Acad Sci U S A 96(10):5482-5485.

Lopez G, Rojas A, Tress M, Valencia A. 2007. Assessment of predictions submitted for the CASP7 function prediction category. Proteins 69 Suppl 8:165-174.

Lundstrom K. 2005. Structural biology of G protein-coupled receptors. Bioorg Med Chem Lett 15(16):3654-3657.

Mac TT, von Hacht A, Hung KC, Dutton RJ, Boyd D, Bardwell JC, Ulmer TS. 2008. Insight into disulfide bond catalysis in Chlamydia from the structure and function of DsbH, a novel oxidoreductase. J Biol Chem 283(2):824-832.

Malmstrom L, Riffle M, Strauss CE, Chivian D, Davis TN, Bonneau R, Baker D. 2007. Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. PLoS Biol 5(4):e76.

Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z and others. 2005. CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res 33(Database issue):D192-196.

Marti-Renom M, Stuart A, Fiser A, Sanchez R, Melo F, Sali A. 2000. Comparative protein structure modeling of genes and genomes Annual Review of Biophysics and Biomolecular Structure 29:291-325.

McGuffin L, Jones D. 2003. Improvement of GenTHREADER method for genomic fold recognition. Bioinformatics 19(7):874-881.

Miao Z, Luker KE, Summers BC, Berahovich R, Bhojani MS, Rehemtulla A, Kleer CG, Essner JJ, Nasevicius A, Luker GD and others. 2007. CXCR7 (RDC1) promotes breast and lung tumor growth in vivo and is expressed on tumor-associated vasculature. Proc Natl Acad Sci U S A 104(40):15735-15740.

Michino M, Abola E, Brooks III CL, Dixon JS, Moult J, Stevens RC. 2009. Community-wide blind assessment of methods for GPCR structure modeling and docking. Nature Reviews: Drug Discovery:Submitted.

Needleman S, Wunsch C. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48:443-453.

Oldziej S, Czaplewski C, Liwo A, Chinchio M, Nanias M, Vila JA, Khalili M, Arnautova YA, Jagielska A, Makowski M and others. 2005. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. Proc Natl Acad Sci U S A 102(21):7547-7552.

Ostermeier C, Michel H. 1997. Crystallization of membrane proteins. Curr Opin Struct Biol 7(5):697-701.

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE and others. 2000. Crystal structure of rhodopsin: A G protein-coupled receptor. Science 289(5480):739-745.

Pisarska M, Mulchahey JJ, Sheriff S, Geracioti TD, Kasckow JW. 2001. Regulation of corticotropin-releasing hormone in vitro. Peptides 22(5):705-712.

Rasmussen SG, Choi HJ, Rosenbaum DM, Kobilka TS, Thian FS, Edwards PC, Burghammer M, Ratnala VR, Sanishvili R, Fischetti RF and others. 2007. Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. Nature 450(7168):383-387.

Read RJ, Chavali G. 2007. Assessment of CASP7 predictions in the high accuracy template-based modeling category. Proteins 69 Suppl 8:27-37.

Rosenbaum DM, Cherezov V, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Yao XJ, Weis WI, Stevens RC and others. 2007. GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. Science 318(5854):1266-1273.

Rost B. 1999. Twilight zone of protein sequence alignments. Protein Eng 12(2):85-94.

Roy A, Kucukural A, Mukherjee S, Hefty PS, Zhang Y. 2010. Large scale benchmarking of protein function prediction using modeled protein structures. J Mol Biol(Submitted).

Sali A, Blundell T. 1993. Comparative protein modelling by satisfaction of spatial restraints. Journal of Molecular Biology 234:779-815.

Sanchez R, Pieper U, Mirkovic N, Bakker Pd, Wittenstein E, Sali A. 2000. MODBASE, a database of annotated comparitive protein structure models Nucleic Acids Research 28(1):250-253.

Sanchez R, Sali A. 1997. Evaluation of comparative protein structure modelling by MODELLER-3. Proteins Suppl 1:50-58.

Sanchez R, Sali A. 1998. Large scale structure modelling of the Saccharomyces cerevisiae genome. Proceedings of The National Academy of Science 95:13597-13602.

Sautel M, Rudolf K, Wittneben H, Herzog H, Martinez R, Munoz M, Eberlein W, Engel W, Walker P, Beck-Sickinger AG. 1996. Neuropeptide Y and the nonpeptide antagonist BIBP 3226 share an overlapping binding site at the human Y1 receptor. Mol Pharmacol 50(2):285-292.

Schwartz TW. 1994. Locating ligand-binding sites in 7TM receptors by protein engineering. Curr Opin Biotechnol 5(4):434-444.

Shi J, Blundell TL, Mizuguchi K. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310(1):243-257.

Shi L, Javitch JA. 2002. The binding site of aminergic G protein-coupled receptors: the transmembrane segments and second extracellular loop. Annu Rev Pharmacol Toxicol 42:437-467.

Simons KT, Kooperberg C, Huang E, Baker D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268(1):209-225.

Simons KT, Strauss C, Baker D. 2001. Prospects for ab initio protein structural genomics. J Mol Biol 306(5):1191-1199.

Sippl M, Weitckus S. 1992. Detection of native like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations Proteins 13:258-271.

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

Skolnick J, Kihara D. 2001. Defrosting the frozen approximation: PROSPECTOR - a new approach to threading. Proteins:Structure,Function and Genetics 42:319-331.

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

Skolnick J, Kihara D, Zhang Y. 2004b. Development and large scale benchmrk testing of the Prospector_3 threading algorithm. Proteins 56(3):502-518.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. J Mol Biol 147(1):195-197.

Soding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951-960.

Tramontano A, Morea V. 2003. Assesment of homology based predictions in CASP 5. Proteins 53(Suppl 6):352-368.

Vitkup D, Melamud E, Moult J, Sander C. 2001. Completeness in structural genomics. Nat Struct Biol 8(6):559-566.

Wallace AC, Laskowski RA, Thornton JM. 1996. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci 5(6):1001-1013.

Warne T, Serrano-Vega MJ, Baker JG, Moukhametzianov R, Edwards PC, Henderson R, Leslie AG, Tate CG, Schertler GF. 2008. Structure of a beta1-adrenergic G-protein-coupled receptor. Nature 454(7203):486-491.

Watson S, Arkinstall S. 1994. The G protein Linked Receptors Factbook. New York: Academic Press.

Wiley SR. 1998. Genomics in the real world. Curr Pharm Des 4(5):417-422.

Wu S, Skolnick J, Zhang Y. 2007a. Ab initio modelling of small proteins by iterative TASSER simulations. BMC Biol 5:17.

Wu S, Zhang Y. 2007b. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res 35(10):3375-3382.

Wu S, Zhang Y. 2008a. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics 24(7):924-931.

Wu S, Zhang Y. 2008b. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72:547-556.

Wu S, Zhang Y. 2009. Improving protein tertiary structure assembly by sequence based contact predictions. Submitted.

Xu Y, Xu D. 2000. Protein threading using PROSPECT: design and evaluation. Proteins 40(3):343-354.

Zhang B, Jaroszewski L, Rychlewski L, Godzik A. 1997. Similarities and differences between non-homologous proteins with similar folds: evaluation of threading strategies. Folding and Design 2(5):307-317.

Zhang Y. 2007. Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 69 Suppl 8:108-117.

Zhang Y. 2008a. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40.

Zhang Y. 2008b. Progress and challenges in protein structure prediction. Curr Opin Struct Biol 18(3):342-348.

Zhang Y. 2009a. I-TASSER: Fully automated protein structure prediction in CASP8. Proteins:In press.

Zhang Y. 2009b. Protein structure prediction: when is it useful? Curr Opin Struct Biol 19(2):145-155.

Zhang Y, Devries ME, Skolnick J. 2006a. Structure modeling of all identified G protein-coupled receptors in the human genome. PLoS Comput Biol 2(2):e13.

Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. 2006b. On the origin and highly likely completeness of single-domain protein structures. Proc Natl Acad Sci U S A 103(8):2605-2610.

Zhang Y, Kihara D, Skolnick J. 2002. Local energy landscape flattening: Parallel hyperbolic monte-carlo sampling of protein folding. Proteins 48:192-201.

Zhang Y, Kolinski A, Skolnick J. 2003. TOUCHSTONE II: a new approach to ab initio protein structure prediction Biophysical Journal 85:1145-1164.

Zhang Y, Skolnick J. 2004a. Automated Structure Prediction of Weekly Homologous Proteins on a Genomic Scale Proceedings of The National Academy of Science 101:7594-7599.

Zhang Y, Skolnick J. 2004b. Spicker: Approach to clustering protein structures for near native model selection. J . of Comp. Chem. 25:865-871.

Zhang Y, Skolnick J. 2004c. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. Biophysical Journal 87:2647-2655.

Zhang Y, Skolnick J. 2005a. The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci U S A 102(4):1029-1034.

Zhang Y, Skolnick J. 2005b. TM-align:a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research 33(7):2302-2309.

Zhou H, Zhou Y. 2004. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 55(4):1005-1013.

Zhou H, Zhou Y. 2005. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58(2):321-328.

Zhou W, Flanagan C, Ballesteros JA, Konvicka K, Davidson JS, Weinstein H, Millar RP, Sealfon SC. 1994. A reciprocal mutation supports helix 2 and helix 7 proximity in the gonadotropin-releasing hormone receptor. Mol Pharmacol 45(2):165-170.