Protein Structure Prediction

Advanced article

Jeffrey Skolnick, *Georgia Institute of Technology, Atlanta, Georgia, USA*

The state of the art of the field of protein structure prediction is reviewed. The strengths and weaknesses of the three general approaches, comparative modelling, threading and template-free modelling, are discussed, and an overview of the results of the critical assessment of structure prediction (CASP) protein structure prediction experiments are summarized. The implications for protein structure prediction of the finding that the library of solved single domain protein structures is likely complete are examined. Recent advances in the modelling of membrane proteins and proteome scale protein structure predictions are presented. Finally, the key remaining unsolved proteins in protein structure prediction are described.

Article Contents

doi: 10.1002/9780470015902.a0003031

Introduction

A paradigm shift brought about by the sequencing of the genomes of hundreds of organisms has occurred in biology. There is increasing focus on the large-scale, high-throughput examination of all genes and gene products of an organism, with the aim of assigning their functions and elucidating how they interact and operate on a system-wide level. This has given rise to the new and emerging field of systems biology. However, having a gene's deoxyribonucleic acid (DNA) sequence does not directly provide insight into its function. Sequence-based methods which detect evolutionary relationships can provide insights into some aspects of the biological function of about 40–60% of the open reading frames (ORFs) found in a given genome but they increasingly fail as the protein families become ever more distant. Predicting the functions of these unassigned ORFs is an important challenge. Because the biochemical function of a protein is ultimately determined by the three-dimensional structure of its biologically active, native conformation, protein structures can assist in functional annotation. The recognition of the role that structure plays in the elucidation of protein function is one impetus for structural genomics that aims for high-throughput protein structure determination. Another is to provide the complete library of solved protein structures so that an arbitrary protein lies within modelling distance of a solved protein structure.

Protein structure prediction methods have historically been divided into three approaches: comparative modelling (CM), threading and template-free (TF) or *ab initio* methods. Conceptually, CM and threading employ the same ideas: (1) Identify a protein that is structurally related to the target sequence of interest. For CM, the target and template proteins are clearly evolutionarily related, while in threading they need not be. Threading should identify homologous as well as analogous folds, viz. proteins that adopt a similar tertiary structure but need not have any evolutionary relationship. (2) Once the related fold is identified, the target sequence is aligned to the template structure either indirectly by performing a sequence alignment and then transferring this alignment to the associated position in the structure as in CM or by incorporating structural information directly into the alignment procedure. With the convergence of threading and CM methods, this has recently given rise to the term template-based (TB) approaches. In TF or *ab initio* methods, one does not use any global template structural information as input to the structure prediction process. Thus, the possibility of assembling a novel fold exists. (As shown later for compact single domain proteins, recent work suggests that there are few if any novel folds (Zhang *et al*., 2006), but the identification of analogous structures remains problematic.) Some approaches are purely physics based and use empirical potentials that are based on quantum mechanics, while others use predicted secondary structure and/or side chain contacts and knowledge-based potentials derived from the properties of solved protein structures.

CASP Evaluation of the State of the Art of Protein Structure Prediction

The status of the field of protein structure prediction has been evaluated on a biannual basis by the critical assessment of structure prediction (CASP) experiments (see later for a more detailed discussion of CASP), where the sequences of structures that are about to be experimentally determined are made available to the protein structure prediction community, the structures are then blindly predicted and then assessed. This offers the advantage that the various approaches can be compared for the same set of targets. However, the number of targets is small, and especially for the TF category, only a handful are typically present. Thus, it is difficult to assess progress, so caution in over interpreting the results should be exercised. Nevertheless, there have been a few general trends. In recent CASPs, the most successful have been unified approaches that span the CM to TF range. ROSETTA (Chivian *et al*., 2005) and TASSER (Zhang *et al*., 2005), were examples of two successful unified approaches in CASP6 (Zhang *et al*., 2005). ROSETTA models the protein using a library of preselected fragments that are three and nine residues in length. TASSER first does threading which provides a set of continuous fragments and predicted tertiary contacts that are then assembled into global folds with the unaligned regions predicted by an *ab initio* approach. Despite this methodological unification, there is a conceptual advantage in discussing the three approaches separately, as they have different success rates and applicability, depending on the target's relationship to proteins whose structures are in the Protein Data Bank (PDB). We discuss the different methodologies later as well as highlight the state of the art for the various level of target difficult as assessed by CASP.

Comparative Modelling/Homology Model Building

Comparative modelling can predict the structure of a target protein whose sequence identity is above 30% to a protein sequence having a solved structure, the template. This is the regime where the alignments are stable and an evolutionary relationship between the target and template proteins can be readily established. For proteins with more than 50% sequence identity to their template, CM sometimes yields models with a 0.5–1 Å root mean-square deviation (RMSD), from the native for the resulting backbone atoms. In the 30–50% sequence identity range, the backbone frequently has about 85% of its core within a RMSD of 3.5 Å from native, with errors mainly in the loops. When the sequence identity drops below 30%, the 'twilight' zone, model accuracy sharply decreases, because of the lack of significant template hits and substantial alignment errors.

As shown in CASP6, a limitation of traditional comparative modelling techniques is that the predicted structures are generally closer to the template structure on which they are based rather than to their native conformation, but the more recent CASPR exercise (a variant of CASP where the starting template structure is provided) began to show progress in this direction. One notable exception in CASP6 was TASSER that often generated an improvement over the template alignment for this identity range (Zhang *et al*., 2005). Furthermore, the recently developed TASSER-lite programme (Pandit *et al*., 2006) (TASSER optimized for the CM limit to give rapid results) was applied to 901 single domain proteins, 41–200 residue in length with sequence identities between 35 and 90% to the template. The results are given in **Figure 1a**, where for the same aligned region, the RMSD of the final model to native is shown. For many cases, there is a clear improvement, with a number of models refined from a RMSD of more than 2 Å to structures close to 1 Å. **Figure 1b** shows results from the widely used CM programme MODELLER5, whose predictions basically recapitulate the template alignment.

Threading Approaches to Protein Structure Prediction

Methodology

As mentioned earlier, the goal of threading is to be able to identify analogous as well as homologous folds. Over the years, there have been a variety of attempts to develop purely structure-based approaches (i.e. those that do not directly exploit evolutionary information) and which represent the protein by its Cαs, Cβs, side-chain centres of mass or other reduced descriptors. Because generating an alignment with explicit consideration of pairwise interactions is NP-hard, pair potentials between interacting residues are often either reduced to pseudo one-body interactions or the structure-based component of the energy simply accounts for predicted secondary structure or burial propensities that may be depth-dependent from the protein's surface. In practice, the most successful threading algorithms

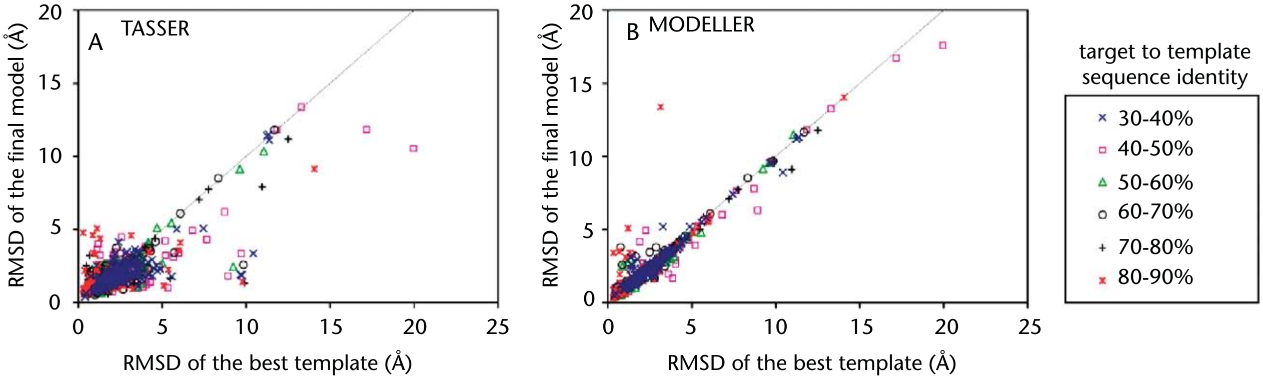

Figure 1 (a) Scatter plot of the RMSD of the final model by TASSER (Zhang and Skolnick, 2004) to native versus RMSD of the initial alignment by the threading algorithm PROSPECTOR_3 (Skolnick *et al*., 2004) to native. The same aligned region is used in both RMSD calculations. (b) Similar to (a), but with models from MODELLER (Pieper *et al*., 2006). The sequence identity ranges are shown in the figure legend.

combine sequence profile information, viz. evolutionary information, with a template structure-based component. Structure-dependent substitution matrices, gap penalties and profiles have also been employed with some success. As first demonstrated in CASP5 and subsequently in CASP6, a number of threading methods significantly outperform PSI-BLAST. By CASP7, the latest variants of these algorithms can sometimes identify structurally related templates at the level of 10% sequence identity with a rather accurate global alignment.

Metapredictor-based protein structure prediction approaches

Recent CASP experiments demonstrated the power of Metapredictors, defined as automated servers that combine structure predictions from a variety of threading and sequence-based servers that make more accurate consensus predictions that now can rival all but the best human predictors. The idea is that for difficult targets, consensus among different methods is more reliable than an individual approach as different threading methods recognize different features, so that in combination better composite results can be obtained (but the results do not improve by much when more than three state-of-the-art methods are used as input). Among Metapredictors of note in CASP6 were 3-D SHOTGUN (Fischer, 2003) which does not just select a model from the input structures but generates more complete and accurate hybrid models by splicing together individual models. In large-scale testing, 3D-SHOTGUN produced models up to 28% more accurate than the individual methods with 17% higher specificity in identifying correct templates.

Structural completeness of the PDB

Essential to the success of CM/threading approaches is the presence of a solved template structure upon which a reasonable model of the target's structure can be assembled. Such templates are what we call '*buildable*'. Is there a limited, but large repertoire of single domain topologies such that at some point, the PDB would be sufficiently complete that the likelihood of finding a new fold is minimal? Or is the repertoire of folds essentially infinite? Consistent with the early enumerations of protein topologies, the preponderance of evidence suggests that the former view holds (Hou *et al.*, 2005). This idea is strongly supported by the recent work of Kim *et al.* who find that the four main classes of protein structures emerge from a common centre 9 with sparse regions in between that possibly arise because certain folds are unstable. If the number of folds is indeed finite, then one can ask how complete is the current PDB library? That is, how likely is it that a given protein will have an already solved structure?

One way to explore this issue is to employ structural alignment algorithms to establish the structural relationship, if any, between a newly solved protein structure and those already in the PDB. (Structural alignment algorithms are designed to define the best structure match between two protein structures when the set of aligned residues are not *a priori* specified.) One can also examine structural relationships between nonhomologous PDB structures as well as between PDB structures and incorrectly folded decoys from *ab initio* folding simulations. In that respect, several authors addressed the nature of protein structure space by comparing all (or representative) structures in the PDB and emphasized its discreteness for protein domains (Hou *et al.*, 2005). However, this conclusion might just reflect the fact that the structural alignment algorithms used lacked the sensitivity to detect more distant structural relationships. Support for the view that structure space is continuous comes from Shindyalov & Bourne who, using their CE method, recently pointed out that substructures obtained from an all-against-all structure comparison sometimes distribute among protein domains, transgressing their respective fold types (Shindyalov and Bourne, 2000). They find ~130 residue long, continuous substructures, much longer than the conventional concept of supersecondary structure. Thus, there are structural motifs of significant length that occur in many other folds, and some regions of protein fold space are not as distinct as once thought. Indeed, in a recent review, Honig *et al.* conclude that structure space is likely continuous and multidimensional (Kolodny *et al.*, 2006).

The continuity of fold space, while suggestive, does not require that the possible fold space of all compact protein structures in the current PDB be complete. Kihara and Skolnick (2003) demonstrated that at the level of single domain proteins, the PDB is likely complete and provides a set of templates on which low-to-moderate resolution structures can be built. More recently, analysing randomly generated, 100 and 200 residues, compact conformations of generic homopolypeptides in simplified and all-atom protein models, Skolnick *et al.* showed that all have similar folds in the PDB, and conversely, all compact, single domain protein structures in the PDB have structural matches to the set of compact homopolypeptide structures (Zhang *et al.*, 2006). Thus, both sets are likely complete, with the protein fold universe arising from compact conformations of hydrogen bonded, secondary structures. Since side chains are represented by the C_{β}s alone in both protein models, these results suggest that the observed protein folds are insensitive to chain packing details. Sequence specificity enters in fine-tuning the structure and stabilizing a given fold with respect to alternatives.

To demonstrate that the resulting set of structures are buildable, the ten worst PDB-compact homopolypeptide structures matched on the basis of their TM-score (a measure of structural similarity that is length independent, ranging from [0,1]). The value is 0.30 for the best structural alignment of a pair randomly related proteins, with a standard deviation of 0.01 and is 1.0 for a pair of identical structures, whose value is about 0.37 were examined. Although this is a modest TM-score, the alignments provide the correct topology for around 2/3 of the core-region of the protein structure. After TASSER refinement, the average TM-score improves to 0.62 (structural alignment Z-score of 32) and the average global RMSD from the PDB structure of the first TASSER model is 5.1 Å. Thus, reasonable full-length models can be built.

Template-free Protein Structure Prediction

The practical issue is to identify a template in the PDB from which a biologically useful model can be built. Unfortunately, for ∼1/3 of compact, nonhomologous single domain proteins such templates cannot be identified using extant threading approaches (Zhang and Skolnick, 2004). Here, TF or *ab initio* methods are required. They are also needed to predict the structure of the unaligned regions/loops and tails in proteins where a reasonable template structure can be identified. Furthermore, the PDB is not likely to be complete at the level of multidomain or multimeric protein structures. TF structure prediction is essential for these situations.

In TF folding, one starts from a random conformation and then attempts to assemble the native structure without use of any template information. Some variants are more physics-based (i.e. they employ little information from native protein structures) and could be used to better understand the factors stabilizing proteins, not to mention to simulate pathways. Recently, there has been progress in the prediction of the tertiary structure of some very small proteins using molecular dynamics simulations of detailed atomic models. Notably, Baker *et al.* generated a model at near atomic resolution in CASP6 (Bradley *et al.*, 2005). However, the CASP6 results suggested that while there are a number of promising developments, the number of TF targets was too small to definitively identify progress (Chivian *et al.*, 2005; Zhang *et al.*, 2005) and that successful TF structure prediction is limited to small, single domain, preferably helical, proteins. As in threading, model ranking can be problematic.

Modelling of Membrane Proteins

Although integral membrane proteins, estimated to comprise 15–40% of the proteome (with the lower (higher) range characteristic of prokaryotes (eukaryotes)) are essential for cellular function, they represent less than 1% of the entire PDB. Thus, the ability to model membrane protein structures is essential. However, because the number of solved structures is so small, CM techniques are of limited validity, and it is difficult to assess the generality of the success of membrane protein structure prediction methods. Since Milik and Skolnick (1992) first successfully modelled the insertion of magainin, M2δ, mellitin, fd and pf1 coat proteins into a membrane by treating the membrane as hydrophobic slab bounded by an interfacial region to mimic the lipid head groups, related simulations have successfully described the tendency of helical membrane peptides to insert and orient with respect to the bilayer (Im and Brooks, 2005). However, modelling individual helical peptides is just a small part of the integral membrane protein tertiary structure prediction problem. Often, to predict the tertiary structure of helical membrane proteins, one first predicts the putative transmembrane region locations in the sequence and then models the packing of the helices and intervening loops.

Since G protein-coupled receptors (GPCRs) are an essential class of integral membrane signalling proteins, they have been the subject of numerous such modelling efforts. The GPCR rhodopsin has received special attention because it has a high-resolution crystal structure. Becker *et al.* used PREDICT to model the transmembrane region with a RMSD from native of 2.9 Å (Becker *et al.*, 2004).Using MembStruk, Vaidehi *et al.* built a model with a RMSD from native of 3.1 Å in the transmembrane region and 8.3 Å for the full-length molecule (Vaidehi *et al.*, 2002). Alternatively, one can use experimental restraints to assist in structure prediction. For example, Sale *et al.* modelled the transmembrane helical region using a statistical potential combined with 27 experimental distance constraints and built a model with a RMSD of 3.2 Å to native in the transmembrane region (Sale *et al.*, 2004). One can also combine modelling with cryo-electron microscopy data. While the results were encouraging, their generality remains to be demonstrated.

Although tertiary structural information is crucial for functional annotation and drug design, there are few experimentally determined GPCR structures. To address this issue, TASSER was employed to generate structure predictions for all 907 putative GPCRs in the human genome (Zhang *et al.*, 2006). First benchmarking of TASSER on membrane proteins with solved structures was done to provide an estimate of the expected success rate. The results for the benchmarking on bovine rhodopsin are summarized later. As shown in **Figure 2**, on excluding homologous structures whose sequence identity is >30% as well as all seven transmembrane helical proteins in the PDB (e.g. bacteriorhodopsin),

PROSPECTOR_3 identified three templates that are quite distant from the rhodopsin structure. After TASSER, a RMSD of 4.6 Å for the final model is obtained on superimposing all 338 C_α atoms (10 residues are absent in the crystal structure). The major errors are in the *N*- and *C*-termini and the C3 loop. Excising the tails and superimposing the model on to the core (residues 32–323), the RMSD to the native structure is 3.3 Å. If we consider the transmembrane helix region, the RMSD is 2.1 Å. Applying the methodology to the four other solved seven transmembrane proteins, archeorhodopsin (1uaz), sensory rhodopsin (1jgj), halorhodopsin (1e12) and bacteriorhodopsin (1ap9), yields final models with RMSDs to native of 2.66, 1.25, 2.39 and 1.86 Å, respectively.

Applying TASSER to a benchmark set of 38 membrane proteins (Zhang *et al.*, 2006) and again excluding templates with >30% sequence identity in the aligned region, 17/38

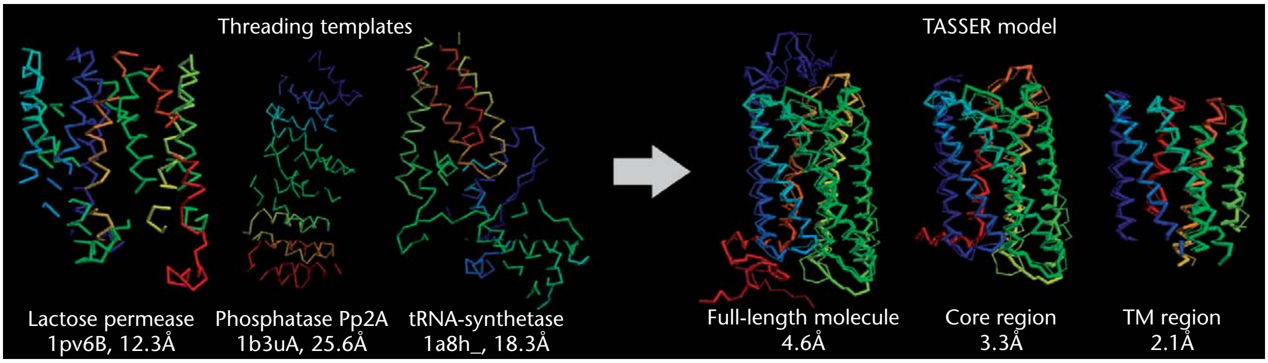

Figure 2 Initial templates from PROSPECTOR_3 and the final TASSER model of highest cluster density superposed on the bovine Rhodopsin crystal structure (PDB ID: 1f88A). Blue to red runs from *N*- to *C*-terminus. The numbers are the root mean-square deviation (RMSD), to native. Here, TM refers to the transmembrane helix region.

(45%) of the targets have a RMSD to native <6.5 Å, and an average RMSD improvement over the template alignment of 4.9 Å. Ninety-seven percent of the targets show an improvement compared to the starting template. While the overall success rate is lower than for water-soluble proteins (66%), it was sufficiently promising to apply TASSER to the human GPCRs.

Based on the distribution of confidence scores shown in **Figure 3**, which also shows the results for the large representative benchmark set of proteins in the PDB below 200 residues, PDB200 benchmark, 820 targets should have the correct fold (Zhang *et al.*, 2006). Models of representative GPCRs were compared with mutagenesis and affinity labelling data, with consistent agreement found. Structure clustering of the predicted models shows that GPCRs with similar structures tend to belong to a similar functional class even when their sequences are diverse. These results demonstrate the usefulness of the *in silico* models for GPCR functional analysis. All predicted GPCR models are available for noncommercial users at http://cssb.biology.gatech.edu/skolnick/files/gpcr/gpcr.html.

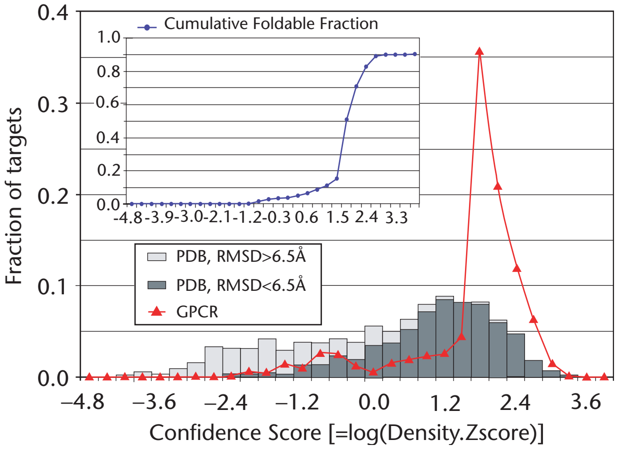

Figure 3 C-score distribution of the predicted models for the 907 human GPCR sequences. The C-score histogram for the PDB200 benchmark is also shown, where dark grey denotes those models with a RMSD <6.5 Å to native and light grey those models whose RMSD is >6.5 Å (Zhang *et al.*, (2006)). Inset: The cumulative foldable fraction calculated assuming that GPCRs have the same correlation between success and C-score as the PDB200 set.

Proteome Scale Protein Structure Predictions

Depending on the proteome, about 40–50% of all sequences have a homologous protein of known structure, with CM results compiled in a number of databases. For example, MODBASE (Pieper *et al.*, 2006) contains over 3 million models for domains covering 60% of the sequences in the UniProt database (Wu *et al.*, 2006). Other CM modelling databases include PEDANT (Riley *et al.*, 2005) that contains structural predictions for 270 bacterial, 23 archaeal and 1 eukaryotic proteomes. Turning to more powerful threading algorithms that can provide approximate structures for proteins in the twilight zone of sequence identity, Kihara and Skolnick showed that threading can provide at least approximate models for about 72% of microbial sequences in representative proteomes (Kihara and Skolnick, 2004). A more comprehensive, threading-based structural database is GENTHREADER (McGuffin *et al.*, 2004) that contains structure predictions for 261 proteomes; on average, the fold coverage statistics for reliable models are consistent with that found by Kihara and Skolnick (2004).

The small proteins in *Mycoplasma genitalium* were the subjects of one of the earliest proteome scale fold prediction studies, where a combined approach that spans the range from CM to TF was applied with a success rate estimated at 60% (Kihara *et al.*, 2002). At about the same time, Baker *et al.* applied ROSETTA to the major Pfam families (Finn *et al.*, 2006) <150 residues in length, with an estimated success rate of about 33%. TASSER was subsequently applied to ORFs <200 residues in the *Escherichia coli* proteome, where 920/1360 ORFs are expected to be accurately predicted based on confidence criterion established in a comprehensive benchmark (see also Figure 3). Finally, as mentioned earlier, TASSER was applied to model all identified human GPCRs with an estimated success rate in predicting biologically useful GPCR models of about 90% (Zhang *et al*., 2006).

The Key Unsolved Problems

To exploit the information provided by the availability of a large number of sequenced genomes, the identification of the function(s) of all genes/gene products in a given organism and the functional relationship among genes in different organisms is essential. Since sequence-based approaches leave the function of about 40–60% of the ORFs of a given genome unassigned, the need for tools that go beyond these methods is acute. With the ongoing structural genomics projects as well as the finding that the PDB is essentially complete for single domain proteins at the level of moderate resolution protein structures, it is apparent that protein structures can make an important contribution to this goal. However, for about 1/3 of single domain proteins that are at best weakly homologous to proteins of solved structure, because suitable templates cannot be identified, extant structure prediction methods do not even yield acceptable low-resolution models. Furthermore, since there are relatively few solved integral membrane protein structures, the development of predictive approaches that can provide even low-to-moderate resolution structures would be of great utility. Moreover, most prediction methods, at best, only implicitly account for the effect of prosthetic groups, ligands and metal ions on protein structure; thus, the development of algorithms to predict the structural differences between the apo and holo forms of a protein is essential. This is tied to the structure prediction of multiple domain proteins, especially when the mutual orientation of the domains depends on whether or not a small molecule ligand is bound. Finally, the ability to refine models close to atomic resolution so that they can be used for ligand screening in drug discovery and in molecular replacement for the solution of crystal structures by X-ray diffraction has not been consistently demonstrated. This inability suggests the need to build better detailed atomic force fields. Thus, while a number of promising approaches to protein structure prediction have been developed, further progress requires that these essential issues be addressed.

Acknowledgements

This work was supported in part by NIH grant Nos. GM-37408, GM-48835 and RR-12255/MMTSB.

References

Becker OM, Marantz Y, Shacham S *et al*. (2004) G protein-coupled receptors: *in silico* drug discovery in 3D. *Proceedings of the National Academy of Sciences of the USA* **101**: 11304–11309.

Bradley P, Misura KM and Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* **309**: 1868–1871.

Chivian D, Kim DE, Malmstrom L *et al*. (2005) Prediction of CASP6 structures using automated Robetta protocols. *Proteins* **61**(suppl. 7): 157–166.

Finn RD, Mistry J, Schuster-Bockler B *et al*. (2006) Pfam: clans, web tools and services. *Nucleic Acids Research* **34**: D247–D251.

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

Hou J, Jun SR, Zhang C and Kim SH (2005) Global mapping of the protein structure space and application in structure-based inference of protein function. *Proceedings of the National Academy of Sciences of the USA* **102**: 3651–3656.

Im W and Brooks CL 3rd (2005) Interfacial folding and membrane insertion of designed peptides studied by molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the USA* **102**: 6771–6776.

Kihara D and Skolnick J (2003) The PDB is a covering set of small protein structures. *Journal of Molecular Biology* **334**: 793–802.

Kihara D and Skolnick J (2004) Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR Q. *Proteins* **55**: 464–473.

Kihara D, Zhang Y, Lu H, Kolinski A and Skolnick J (2002) *Ab initio* protein structure prediction on a genomic scale: application to the *Mycoplasma genitalium* genome. *Proceedings of the National Academy of Sciences of the USA* **99**: 5993–5998.

Kolodny R, Petrey D and Honig B (2006) Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Current Opinion in Structural Biology* **16**: 393–398.

McGuffin LJ, Street S, Sorensen SA and Jones DT (2004) The genomic threading database. *Bioinformatics* **20**: 131–132.

Milik M and Skolnick J (1992) Spontaneous insertion of polypeptide chains into membranes: a Monte Carlo model. *Proceedings of the National Academy of Sciences of the USA* **89**: 9391–9395.

Pandit SB, Zhang Y and Skolnick J (2006) TASSER-Lite: an automated tool for protein comparative modeling. *Biophysical Journal* **91**: 4180–4190.

Pieper U, Eswar N, Davis FP *et al*. (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research* **34**: D291–D295.

Riley ML, Schmidt T, Wagner C, Mewes HW and Frishman D (2005) The PEDANT genome database in 2005. *Nucleic Acids Research* **33**: D308–D310.

Sale K, Faulon JL, Gray GA, Schoeniger JS and Young MM (2004) Optimal bundling of transmembrane helices using sparse distance constraints. *Protein Science* **13**: 2613–2627.

Shindyalov IN and Bourne PE (2000) An alternative view of protein fold space. *Proteins* **38**: 247–260.

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

Vaidehi N, Floriano WB, Trabanino R *et al*. (2002) Prediction of structure and function of G protein-coupled receptors.

Proceedings of the National Academy of Sciences of the USA **99**: 12622–12627.

Wu CH, Apweiler R, Bairoch A *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research* **34**: D187–D191.

Zhang Y, Arakaki AK and Skolnick J (2005) TASSER: an automated method for the prediction of protein tertiary structures in CASP. *Proteins* **61**(suppl. 7): 91–98.

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

Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E and Skolnick J (2006) On the origin and highly likely completeness of single-domain protein structures. *Proceedings of the National Academy of Sciences of the USA* **103**: 2605–2610.

Zhang Y and Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences of the USA* **101**: 7594–7599.

Further Reading

Aloy P and Russell RB (2004) Ten thousand interactions for the molecular biologist. *Nature Biotechnology* **22**: 1317–1321.

Baker D and Sali A (2001) Protein structure prediction and structural genomics. *Science* **294**: 93–96.

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

Chothia C and Finkelstein AV (1990) The classification and origins of protein folding patterns. *Annual Review of Biochemistry* **59**: 1007–1039.

John B and Sali A (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Research* **31**: 3982–3992.

Khalili M, Liwo A and Scheraga HA (2006) Kinetic studies of folding of the B-domain of staphylococcal protein A with molecular dynamics and a united-residue (UNRES) model of polypeptide chains. *Journal of Molecular Biology* **355**: 536–547.

Lee MR, Tsai J, Baker D and Kollman PA (2001) Molecular dynamics in the endgame of protein structure prediction. *Journal of Molecular Biology* **313**: 417–430.

Lundstrom J, Rychlewski L, Bujnicki J and Elofsson A (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Science* **10**: 2354–2362.

Miyazawa S and Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology* **256**: 623–644.

Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E and Skolnick J (2006) On the origin and highly likely completeness of single-domain protein structures. *Proceedings of the National Academy of Sciences of the USA* **103**: 2605–2610.

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