4.10 Comparative Modeling of Drug Target Proteins

N Eswar and A Sali, University of California at San Francisco, San Francisco, CA, USA

4.10.1 Introduction

4.10.1.1 Structure-Based Drug Discovery

Over the past few years, structure-based or rational drug discovery has resulted in a number of drugs on the market and many more in the development pipeline.[1–4] Structure-based methods are now routinely used in almost all stages of drug development, from target identification to lead optimization.[5–8] Central to all structure-based discovery approaches is the knowledge of the three-dimensional (3D) structure of the target protein or complex because the structure and dynamics of the target determine which ligands it binds. The 3D structures of the target proteins are best determined by experimental methods that yield solutions at atomic resolution, such as x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy.[9] Recent developments in the techniques of experimental structure determination have enhanced the applicability, accuracy, and speed of structural studies.[10,11] Despite these advances, however, structural characterization of sequences remains an expensive and time-consuming task.

4.10.1.2 The Sequence–Structure Gap

The publicly available Protein Data Bank (PDB)[12] currently contains ~33 000 structures and grows at a rate of approximately 40% every 2 years. On the other hand, the various genome-sequencing projects have resulted in ~2.1 million sequences, including the complete genetic blueprints of humans and hundreds of other organisms.[13,14] This achievement has resulted in a vast collection of sequence information about possible target proteins with little or no structural information. Current statistics show that the structures available in the PDB account for only ~1.5% of the sequences in the UniProt database.[13] Moreover, the rate of growth of the sequence information is more than twice that of the structures. Due to this wide sequence–structure gap, reliance on experimentally determined structures limits the number of proteins that can be targeted by structure-based drug discovery.

4.10.1.3 Structure Prediction Addresses the Sequence–Structure Gap

Fortunately, domains in protein sequences are gradually evolving entities that can be clustered into a relatively small number of families with similar sequences and structures.[15,16] For instance, 75–80% of the sequences in the UniProt database have been grouped into fewer than 15 000 domain families.[17,18] Similarly, all the structures in the PDB have been classified into about 1000 distinct folds.[19,20] Computational protein structure prediction methods, such as threading[21] and comparative protein structure modeling,[22,23] strive to bridge the sequence–structure gap by utilizing these evolutionary relationships. The speed, low cost, and relative accuracy of these computational methods have led to the use of predicted 3D structures in the drug discovery process.[24,25] The other class of prediction methods, de novo or ab initio methods, attempts to predict the structure from sequence alone, without reliance on evolutionary relationships. However, despite recent progress in these methods,[26] especially for small proteins with fewer than 100 amino acid residues, comparative modeling remains the most reliable method of predicting the 3D structure of a protein, with an accuracy that can be comparable to a low-resolution, experimentally determined structure.[9]

4.10.1.4 The Basis of Comparative Modeling

The primary requirement for reliable comparative modeling is a detectable similarity between the sequence of interest (target sequence) and a known structure (template). As early as 1986, Chothia and Lesk[27] showed that there is a strong correlation between sequence and structural similarities. This correlation provides the basis of comparative modeling, allows a coarse assessment of model errors, and also highlights one of its major challenges: modeling the structural differences between the template and target structures[28] (**Figure 1**).

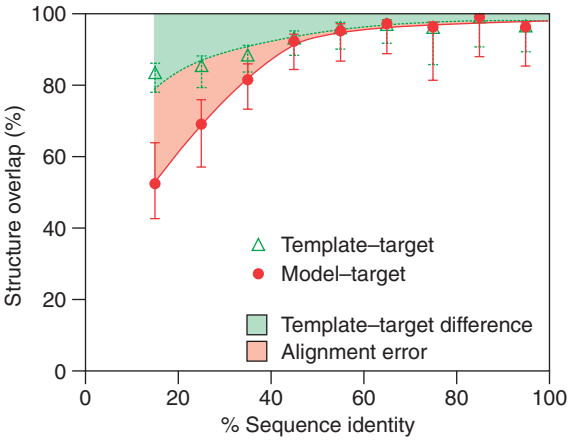

Figure 1 Average model accuracy as a function of sequence identity.[28] As the sequence identity between the target sequence and the template structure decreases, the average structural similarity between the template and the target also decreases (dashed line, triangles).[27] Structural overlap is defined as the fraction of equivalent C^{α} atoms. For the comparison of the model with the actual structure (filled circles), two C^{α} atoms were considered equivalent if they belonged to the same residue and were within 3.5 Å of each other after least-squares superposition. For comparisons between the template structure and the actual target structure (triangles), two C^{α} atoms were considered equivalent if they were within 3.5 Å of each other after alignment and rigid-body superposition. The difference between the model and the actual target structure is a combination of the target–template differences (green area) and the alignment errors (red area). The figure was constructed by calculating 3993 comparative models based on a single template of varying similarity to the targets. All targets had known (experimentally determined) structures.[28]

4.10.1.5 Comparative Modeling Benefits from Structural Genomics

Comparative modeling stands to benefit greatly from the structural genomics initiative.[29] Structural genomics aims to achieve significant structural coverage of the sequence space with an efficient combination of experimental and prediction methods.[30] This goal is pursued by careful selection of target proteins for structure determination by x-ray crystallography and NMR spectroscopy, such that most other sequences are within 'modeling distance' (e.g., >30% sequence identity) of a known structure.[15,16,29,31] The expectation is that the determination of these structures combined with comparative modeling will yield useful structural information for the largest possible fraction of sequences in the shortest possible timeframe. The impact of structural genomics is illustrated by comparative modeling based on the structures determined by the New York Structural Genomics Research Consortium. For each new structure, on average, 100 protein sequences without any prior structural characterization could be modeled at least at the level of the fold.[32] Thus, the structures of most proteins will eventually be predicted by computation, not determined by experiment.

4.10.1.6 Outline

In this review, we begin by describing the various steps involved in comparative modeling. Next, we emphasize two aspects of model refinement, loop modeling and side-chain modeling, due to their relevance in ligand docking and rational drug discovery. We then discuss the errors in comparative models. Finally, we describe the role of comparative modeling in drug discovery, focusing on ligand docking against comparative models. We compare successes of docking against models and x-ray structures, and illustrate the computational docking against models with a number of examples. We conclude with a summary of topics that will impact on the future utility of comparative modeling in drug discovery, including an automation and integration of resources required for comparative modeling and ligand docking.

4.10.2 Steps in Comparative Modeling

Comparative modeling consists of four main steps[23] (**Figure 2a**): (1) fold assignment that identifies similarity between the target sequence of interest and at least one known protein structure (the template); (2) alignment of the target sequence and the template(s); (3) building a model based on the alignment with the chosen template(s); and (4) predicting model errors.

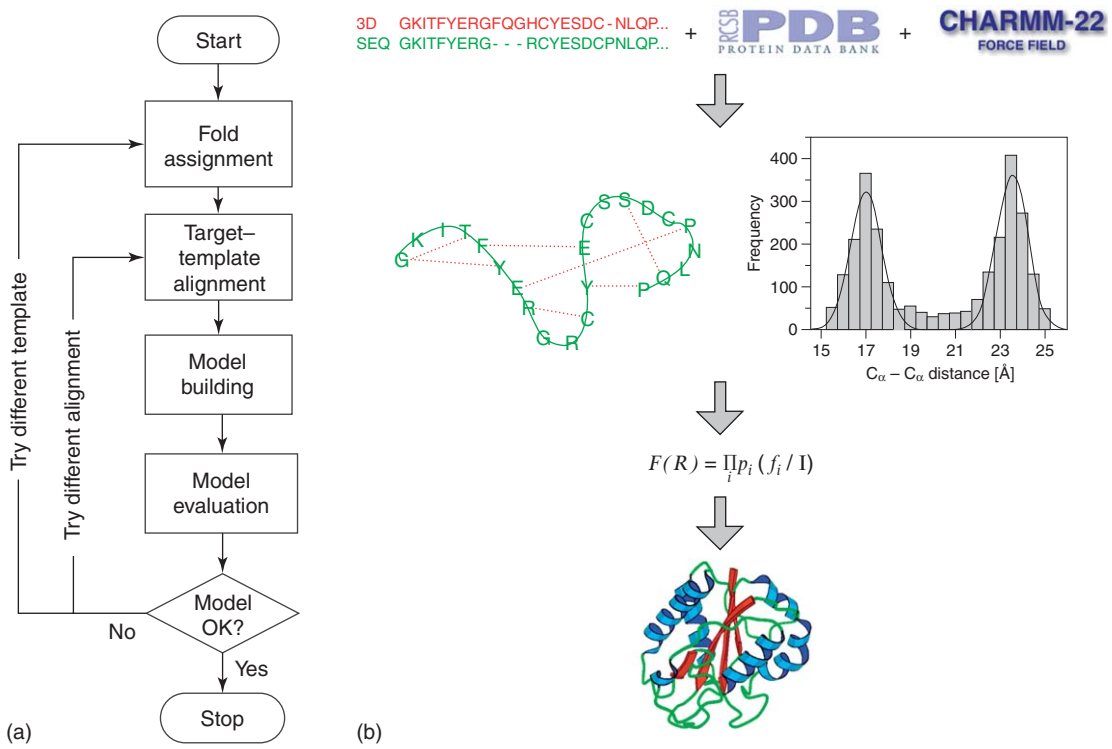

Figure 2 Comparative protein structure modeling. (a) A flowchart illustrating the steps in the construction of a comparative model.[23] (b) Description of comparative modeling by extraction of spatial restraints as implemented in MODELLER.[96] By default, spatial restraints in MODELLER involve: (1) homology-derived restraints from the aligned template structures; (2) statistical restraints derived from all known protein structures; and (3) stereochemical restraints from the CHARMM-22 molecular mechanics force field. These restraints are combined into an objective function that is then optimized to calculate the final 3D structure of the target sequence.

4.10.2.1 Fold Assignment and Sequence–Structure Alignment

Although fold assignment and sequence–structure alignment are logically two distinct steps in the process of comparative modeling, in practice almost all fold assignment methods also provide sequence–structure alignments. In the past, fold assignment methods were optimized for better sensitivity in detecting remotely related homologs, often at the cost of alignment accuracy. However, recent methods simultaneously optimize both the sensitivity and alignment accuracy. Therefore, in the following discussion, we will treat fold assignment and sequence–structure alignment as a single protocol, explaining the differences as needed.

4.10.2.1.1 Fold assignment

As mentioned earlier, the primary requirement for comparative modeling is the identification of one or more known template structures with detectable similarity to the target sequence. The identification of suitable templates is achieved by scanning structure databases, such as PDB,[12] SCOP,[19] DALI,[33] and CATH,[20] with the target sequence as the query. The detected similarity is usually quantified in terms of sequence identity or statistical measures, such as *E*-value or *z*-score, depending on the method used.

4.10.2.1.2 Three levels of similarity

Sequence–structure relationships are coarsely classified into three different regimes in the sequence similarity spectrum: (1) the easily detected relationships characterized by >30% sequence identity; (2) the 'twilight zone,'[34] corresponding to relationships with statistically significant sequence similarity in the 10–30% range; and (3) the 'midnight zone,'[34] corresponding to statistically insignificant sequence similarity.

4.10.2.1.3 Sequence–sequence methods

For closely related protein sequences with identities higher than 30–40%, the alignments produced by all methods are almost always largely correct. The quickest way to search for suitable templates in this regime is to use simple pairwise

sequence alignment methods such as SSEARCH,[35] BLAST,[36] and FASTA.[35] Brenner *et al.* showed that these methods detect only ~18% of the homologous pairs at less than 40% sequence identity, while they identify more than 90% of the relationships when sequence identity is between 30% and 40%.[37] Another benchmark, based on 200 reference structural alignments with 0–40% sequence identity, indicated that BLAST is able to correctly align only 26% of the residue positions.[46]

4.10.2.1.4 Sequence–profile methods

The sensitivity of the search and accuracy of the alignment become progressively difficult as the relationships move into the twilight zone.[34,38] A significant improvement in this area was the introduction of profile methods by Gribskov and co-workers.[39] The profile of a sequence is derived from a multiple sequence alignment and specifies residue-type occurrences for each alignment position. The information in a multiple sequence alignment is most often encoded as either a position-specific scoring matrix (PSSM)[36,40,41] or as a hidden Markov model (HMM).[42,43] In order to identify suitable templates for comparative modeling, the profile of the target sequence is used to search against a database of template sequences. The profile–sequence methods are more sensitive in detecting related structures in the twilight zone than the pairwise sequence-based methods; they detect approximately twice the number of homologs under 40% sequence identity.[44–46] The resulting profile–sequence alignments correctly align approximately 43–48% of residues in the 0–40% sequence identity range[46,47]; this number is almost twice as large as that of the pairwise sequence methods. Frequently used programs for profile–sequence alignment are PSI-BLAST,[36] SAM,[48] HMMER,[42] and BUILD_PROFILE.[49]

4.10.2.1.5 Profile–profile methods

As a natural extension, the profile–sequence alignment methods have led to profile–profile alignment methods that search for suitable template structures by scanning the profile of the target sequence against a database of template profiles, as opposed to a database of template sequences. These methods have proven to include the most sensitive and accurate fold assignment and alignment protocols to date.[47,50–52] Profile–profile methods detect ~28% more relationships at the superfamily level and improve the alignment accuracy by 15–20% compared to profile–sequence methods.[47,53] There are a number of variants of profile–profile alignment methods that differ in the scoring functions they use.[47,50,53–59] However, several analyses have shown that the overall performances of these methods are comparable.[47,50–52] Some of the programs that can be used to detect suitable templates are FFAS,[60] SP3,[53] SALIGN,[47] and PPSCAN.[49]

4.10.2.1.6 Sequence–structure threading methods

As the sequence identity drops below the threshold of the twilight zone, there is usually insufficient signal in the sequences or their profiles for the sequence-based methods discussed above to detect true relationships.[44] Sequence–structure threading methods are most useful in this regime as they can sometimes recognize common folds, even in the absence of any statistically significant sequence similarity.[21] These methods achieve higher sensitivity by using structural information derived from the templates. The accuracy of a sequence–structure match is assessed by the score of a corresponding coarse model and not by sequence similarity, as in sequence comparison methods.[21] The scoring scheme used to evaluate the accuracy is either based on residue substitution tables dependent on structural features such as solvent exposure, secondary structure type, and hydrogen bonding properties,[53,61–63] or on statistical potentials for residue interactions implied by the alignment.[64–68] The use of structural data does not have to be restricted to the structure side of the aligned sequence–structure pair. For example, SAM-T02 makes use of the predicted local structure for the target sequence to enhance homolog detection and alignment accuracy.[69] Commonly used threading programs are GenTHREADER,[61,70] 3D-PSSM,[71] FUGUE,[63] SP3,[53] and SAM-T02 multitrack HMM.[62,69]

4.10.2.1.7 Iterative sequence–structure alignment

Yet another strategy is to optimize the alignment by iterating over the process of calculating alignments, building models, and evaluating models. Such a protocol can sample alignments that are not statistically significant and identify the alignment that yields the best model. Although this procedure can be time-consuming, it can significantly improve the accuracy of the resulting comparative models in difficult cases.[72]

4.10.2.2 Alignment Errors are Unrecoverable

Regardless of the method used, searching in the twilight and midnight zones of the sequence–structure relationship often results in false negatives, false positives, or alignments that contain an increasingly large number of gaps and

alignment errors. Improving the performance and accuracy of methods in this regime remains one of the main tasks of comparative modeling today.[73] It is imperative to calculate an accurate alignment between the target–template pair, as comparative modeling can almost never recover from an alignment error.[74]

4.10.2.3 Template Selection

After a list of all related protein structures and their alignments with the target sequence have been obtained, template structures are prioritized depending on the purpose of the comparative model. Template structures may be chosen purely based on the target–template sequence identity or a combination of several other criteria, such as experimental accuracy of the structures (resolution of x-ray structures, number of restraints per residue for NMR structures), conservation of active-site residues, holo-structures that have bound ligands of interest, and prior biological information that pertains to the solvent, pH, and quaternary contacts. It is not necessary to select only one template. In fact, the use of several templates approximately equidistant from the target sequence generally increases the model accuracy.[75,76]

4.10.3 Model Building

4.10.3.1 Three Approaches to Comparative Model Building

Once an initial target–template alignment is built, a variety of methods can be used to construct a 3D model for the target protein.[23,74,77–80] The original and still widely used method is modeling by rigid-body assembly.[78,79,81] This method constructs the model from a few core regions, and from loops and side chains that are obtained by dissecting related structures. Commonly used programs that implement this method are COMPOSER,[82–85] 3D-JIGSAW,[86] and SWISS-MODEL.[87] Another family of methods, modeling by segment matching, relies on the approximate positions of conserved atoms from the templates to calculate the coordinates of other atoms.[88–92] An instance of this approach is implemented in SegMod.[91] The third group of methods, modeling by satisfaction of spatial restraints, uses either distance geometry or optimization techniques to satisfy spatial restraints obtained from the alignment of the target sequences with the template structures.[93–97] Specifically, MODELLER,[96,98,99] our own program for comparative modeling, belongs to this group of methods.

4.10.3.2 MODELLER: Comparative Modeling by Satisfaction of Spatial Restraints

MODELLER implements comparative protein structure modeling by the satisfaction of spatial restraints that include: (1) homology-derived restraints on the distances and dihedral angles in the target sequence, extracted from its alignment with the template structures[96]; (2) stereochemical restraints such as bond length and bond angle preferences, obtained from the CHARMM-22 molecular mechanics force field[100]; (3) statistical preferences for dihedral angles and nonbonded interatomic distances, obtained from a representative set of known protein structures[101]; and (4) optional manually curated restraints, such as those from NMR spectroscopy, rules of secondary structure packing, cross-linking experiments, fluorescence spectroscopy, image reconstruction from electron microscopy, site-directed mutagenesis, and intuition (**Figure 2b**). The spatial restraints, expressed as probability density functions, are combined into an objective function that is optimized by a combination of conjugate gradients and molecular dynamics with simulated annealing. This model-building procedure is similar to structure determination by NMR spectroscopy.

4.10.3.3 Relative Accuracy, Flexibility, and Automation

Accuracies of the various model-building methods are relatively similar when used optimally.[102,103] Other factors, such as template selection and alignment accuracy, usually have a larger impact on the model accuracy, especially for models based on less than 30% sequence identity to the templates. However, it is important that a modeling method allows a degree of flexibility and automation to obtain better models more easily and rapidly. For example, a method should allow for an easy recalculation of a model when a change is made in the alignment; it should be straightforward to calculate models based on several templates; and the method should provide tools for incorporation of prior knowledge about the target (e.g., cross-linking restraints and predicted secondary structure).

4.10.4 Refinement of Comparative Models

Protein sequences evolve through a series of amino acid residue substitutions, insertions, and deletions. While substitutions can occur throughout the length of the sequence, insertions and deletions mostly occur on the surface of proteins in segments that connect regular secondary structure segments (i.e., loops). While the template structures are helpful in the modeling of the aligned target backbone segments, they are generally less valuable for the modeling of side chains and irrelevant for the modeling of insertions such as loops. The loops and side chains of comparative models are especially important for ligand docking; thus, we discuss them in the following two sections.

4.10.4.1 Loop Modeling

4.10.4.1.1 Definition of the problem

Loop modeling is an especially important aspect of comparative modeling in the range from 30% to 50% sequence identity. In this range of overall similarity, loops among the homologs vary while the core regions are still relatively conserved and aligned accurately. Loops often play an important role in defining the functional specificity of a given protein, forming the active and binding sites. Loop modeling can be seen as a mini protein folding problem because the correct conformation of a given segment of a polypeptide chain has to be calculated mainly from the sequence of the segment itself. However, loops are generally too short to provide sufficient information about their local fold. Even identical decapeptides in different proteins do not always have the same conformation.[104,105] Some additional restraints are provided by the core anchor regions that span the loop and by the structure of the rest of the protein that cradles the loop. Although many loop-modeling methods have been described, it is still challenging to correctly and confidently model loops longer than approximately 8–10 residues.[98,106]

4.10.4.1.2 Two classes of methods

There are two main classes of loop-modeling methods: (1) database search approaches that scan a database of all known protein structures to find segments fitting the anchor core regions[90,107]; and (2) conformational search approaches that rely on optimizing a scoring function.[108–110] There are also methods that combine these two approaches.[111,112]

4.10.4.1.2.1 Database-based loop modeling

The database search approach to loop modeling is accurate and efficient when a database of specific loops is created to address the modeling of the same class of loops, such as β-hairpins,[113] or loops on a specific fold, such as the hypervariable regions in the immunoglobulin fold.[107,114] There are attempts to classify loop conformations into more general categories, thus extending the applicability of the database search approach.[115–117] However, the database methods are limited because the number of possible conformations increases exponentially with the length of a loop. As a result, only loops up to 4–7 residues long have most of their conceivable conformations present in the database of known protein structures.[118,119] This limitation is made even worse by the requirement for an overlap of at least one residue between the database fragment and the anchor core regions, which means that modeling a 5-residue insertion requires at least a 7-residue fragment from the database.[89] Despite the rapid growth of the database of known structures, it does not seem possible to cover most of the conformations of a 9-residue segment in the foreseeable future. On the other hand, most of the insertions in a family of homologous proteins are shorter than 10–12 residues.[98]

4.10.4.1.2.2 Optimization-based methods

To overcome the limitations of the database search methods, conformational search methods were developed.[108,109] There are many such methods, exploiting different protein representations, objective functions, and optimization or enumeration algorithms. The search algorithms include the minimum perturbation method,[120] molecular dynamics simulations,[111,121] genetic algorithms,[122] Monte Carlo and simulated annealing,[123–125] multiple-copy simultaneous search,[126] self-consistent field optimization,[127] and enumeration based on graph theory.[128] The accuracy of loop predictions can be further improved by clustering the sampled loop conformations and partially accounting for the entropic contribution to the free energy.[129] Another way of improving the accuracy of loop predictions is to consider the solvent effects. Improvements in implicit solvation models, such as the Generalized Born solvation model, motivated their use in loop modeling. The solvent contribution to the free energy can be added to the scoring function for optimization, or it can be used to rank the sampled loop conformations after they are generated with a scoring function that does not include the solvent terms.[98,130–132]

4.10.4.2 Side-Chain Modeling

4.10.4.2.1 Fixed backbone

Two simplifications are frequently applied in the modeling of side-chain conformations.[133] First, amino acid residue replacements often leave the backbone structure almost unchanged,[26] allowing us to fix the backbone during the search for the best side-chain conformations. Second, most side chains in high-resolution crystallographic structures can be represented by a limited number of conformers that comply with stereochemical and energetic constraints.[134] This observation motivated Ponder and Richards[135] to develop the first library of side-chain rotamers for the 17 types of residues with dihedral angle degrees of freedom in their side chains, based on 10 high-resolution protein structures determined by x-ray crystallography. Subsequently, a number of additional libraries have been derived.[136–142]

4.10.4.2.2 Rotamers

Rotamers on a fixed backbone are often used when all the side chains need to be modeled on a given backbone. This approach reduces the combinatorial explosion associated with a full conformational search of all the side chains, and is applied by some comparative modeling[78] and protein design approaches.[143] However, ~15% of the side chains cannot be represented well by these libraries.[144] In addition, it has been shown that the accuracy of side-chain modeling on a fixed backbone decreases rapidly when the backbone errors are larger than 0.5 Å.[145]

4.10.4.2.3 Methods

Earlier methods for side-chain modeling often put less emphasis on the energy or scoring function. The function was usually greatly simplified, and consisted of the empirical rotamer preferences and simple repulsion terms for nonbonded contacts.[138] Nevertheless, these approaches have been justified by their performance. For example, a method based on a rotamer library compared favorably with that based on a molecular mechanics force field,[146] and new methods continue to be based on the rotamer library approach.[147,148] The various optimization approaches include a Monte Carlo simulation,[149] simulated annealing,[150] a combination of Monte Carlo and simulated annealing,[151] the dead-end elimination theorem,[152,153] genetic algorithms,[142] neural network with simulated annealing,[154] mean field optimization,[155] and combinatorial searches.[138,156,157] Several recent papers focused on the testing of more sophisticated potential functions for conformational search[157,158] and development of new scoring functions for side-chain modeling,[159] reporting higher accuracy than earlier studies.

4.10.5 Errors in Comparative Models

The major sources of error in comparative modeling are discussed in the relevant sections above. The following is a summary of these errors, dividing them into five categories (**Figure 3**).

4.10.5.1 Selection of Incorrect Templates

This error is a potential problem when distantly related proteins are used as templates (i.e., less than 30% sequence identity). Distinguishing between a model based on an incorrect template and a model based on an incorrect alignment with a correct template is difficult. In both cases, the evaluation methods (below) will predict an unreliable model. The conservation of the key functional or structural residues in the target sequence increases the confidence in a given fold assignment.

4.10.5.2 Errors due to Misalignments

The single source of errors with the largest impact on comparative modeling is misalignments, especially when the target–template sequence identity decreases below 30%. Alignment errors can be minimized in two ways. Using the profile-based methods discussed above usually results in more accurate alignments than those from pairwise sequence alignment methods. Another way of improving the alignment is iteratively to modify those regions in the alignment that correspond to predicted errors in the model.[75]

4.10.5.3 Errors in Regions without a Template

Segments of the target sequence that have no equivalent region in the template structure (i.e., insertions or loops) are one of the most difficult regions to model. Again, when the target and template are distantly related, errors in the alignment can lead to incorrect positions of the insertions. Using alignment methods that incorporate structural

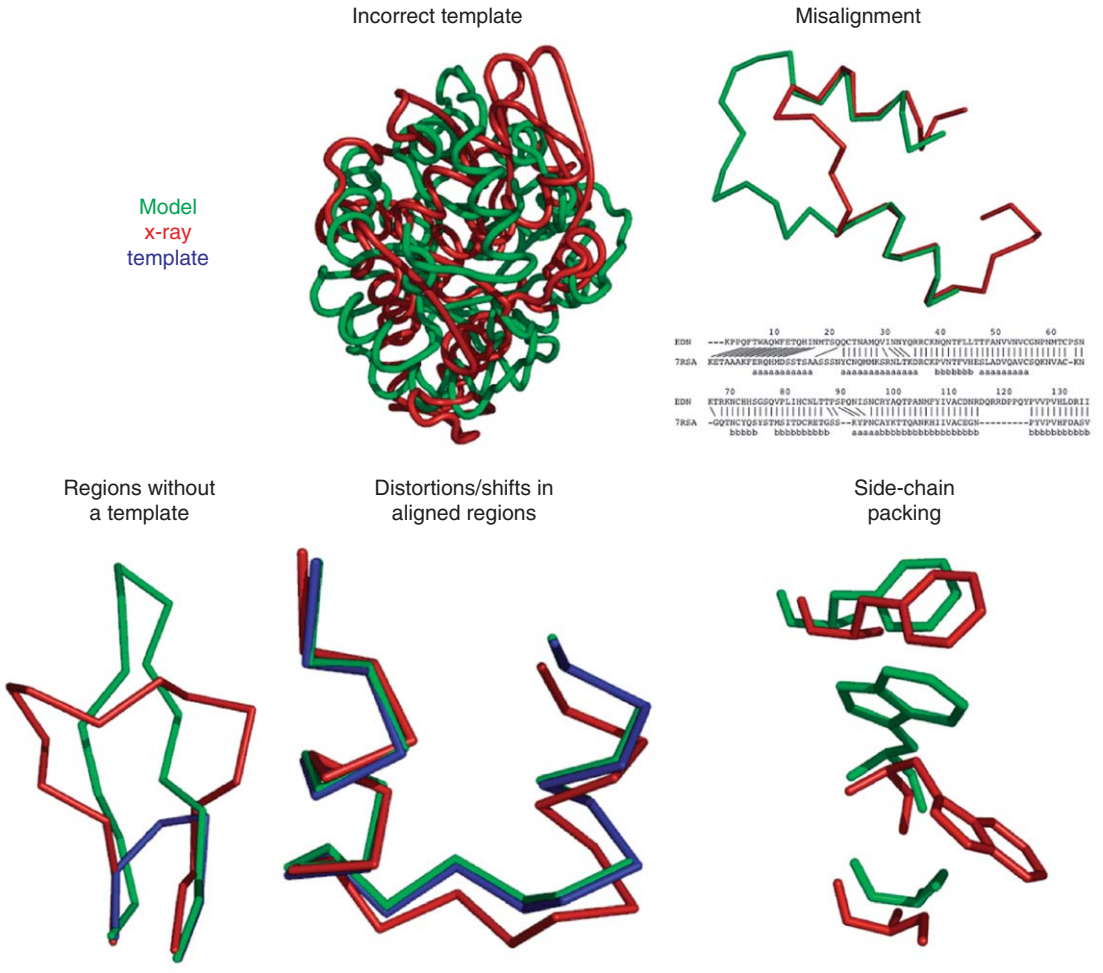

Figure 3 Typical errors in comparative modeling.[23] Shown are the typical sources of errors encountered in comparative models. Two of the major sources of errors in comparative modeling are due to incorrect templates or incorrect alignments with the correct templates. The modeling procedure can rarely recover from such errors. The next significant source of errors arises from regions in the target with no corresponding region in the template, i.e., insertions or loops. Other sources of errors, which occur even with an accurate alignment, are due to rigid-body shifts, distortions in the backbone, and errors in the packing of side chains.

information can often correct such errors. Once a reliable alignment is obtained, various modeling protocols can predict the loop conformation, for insertions of fewer than 8–10 residues.[98,106,111,169]

4.10.5.4 Distortions and Shifts in Correctly Aligned Regions

As a consequence of sequence divergence, the main-chain conformation changes, even if the overall fold remains the same. Therefore, it is possible that in some correctly aligned segments of a model, the template is locally different (<3 Å) from the target, resulting in errors in that region. The structural differences are sometimes not due to differences in sequence, but are a consequence of artifacts in structure determination or structure determination in different environments (e.g., packing of subunits in a crystal). The simultaneous use of several templates can minimize this kind of an error.[75,76]

4.10.5.5 Errors in Side-Chain Packing

As the sequences diverge, the packing of the atoms in the protein core changes. Sometimes even the conformation of identical side chains is not conserved – a pitfall for many comparative modeling methods. Side-chain errors are critical if they occur in regions that are involved in protein function, such as active sites and ligand-binding sites.

4.10.6 Prediction of Model Errors

The accuracy of the predicted model determines the information that can be extracted from it. Thus, estimating the accuracy of a model in the absence of the known structure is essential for interpreting it.

4.10.6.1 Initial Assessment of the Fold

As discussed earlier, a model calculated using a template structure that shares more than 30% sequence identity is indicative of an overall accurate structure. However, when the sequence identity is lower, the first aspect of model evaluation is to confirm whether or not a correct template was used for modeling. It is often the case, when operating in this regime, that the fold assignment step produces only false positives. A further complication is that at such low similarities the alignment generally contains many errors, making it difficult to distinguish between an incorrect template on one hand and an incorrect alignment with a correct template on the other hand. There are several methods that use 3D profiles and statistical potentials,[65,160,161] which assess the compatibility between the sequence and modeled structure by evaluating the environment of each residue in a model with respect to the expected environment, as found in native high-resolution experimental structures. These methods can be used to assess whether or not the correct template was used for the modeling. They include VERIFY3D,[160] PROSAII,[162] HARMONY,[163] ANOLEA,[164] and DFIRE.[165]

Even when the model is based on alignments that have $>30\%$ sequence identity, other factors, including the environment, can strongly influence the accuracy of a model. For instance, some calcium-binding proteins undergo large conformational changes when bound to calcium. If a calcium-free template is used to model the calcium-bound state of the target, it is likely that the model will be incorrect, irrespective of the target–template similarity or accuracy of the template structure.[166]

4.10.6.2 Self-Consistency

The model should also be subjected to evaluations of self-consistency to ensure that it satisfies the restraints used to calculate it. Additionally, the stereochemistry of the model (e.g., bond lengths, bond angles, backbone torsion angles, and nonbonded contacts) may be evaluated using programs such as PROCHECK[167] and WHATCHECK.[168] Although errors in stereochemistry are rare and less informative than errors detected by statistical potentials, a cluster of stereochemical errors may indicate that there are larger errors (e.g., alignment errors) in that region.

4.10.7 Evaluation of Comparative Modeling Methods

It is crucial for method developers and users alike to assess the accuracy of their methods. An attempt to address this problem has been made by the Critical Assessment of Techniques for Proteins Structure Prediction (CASP)[170] and the Critical Assessment of Fully Automated Structure Prediction (CAFASP) experiments.[171] However, both CASP and CAFASP assess methods only over a limited number of target protein sequences.[102,172] To overcome this limitation, two additional evaluation experiments have been described, LiveBench[172] and EVA.[173,174] EVA is a large-scale and continuously running web server that automatically assesses protein structure prediction servers in the categories of secondary structure prediction, residue–residue contact prediction, fold assignment, and comparative modeling. The aims of EVA are: (1) to evaluate continuously and automatically blind predictions by prediction servers, based on identical and sufficiently large data sets; (2) to provide weekly updates of the method assessments on the web; and (3) to enable developers, nonexpert users, and reviewers to determine the performance of the tested prediction servers.

4.10.8 Applications of Comparative Models

There is a wide range of applications of protein structure models (**Figure 4**).[1,175–180] For example, high- and medium-accuracy comparative models are frequently helpful in refining functional predictions that have been based on a sequence match alone because ligand binding is more directly determined by the structure of the binding site than by its sequence. It is often possible to predict correctly features of the target protein that do not occur in the template structure.[181,182] For example, the size of a ligand may be predicted from the volume of the binding site cleft and the location of a binding site for a charged ligand can be predicted from a cluster of charged residues on the protein. Fortunately, errors in the functionally important regions in comparative models are many times relatively low because

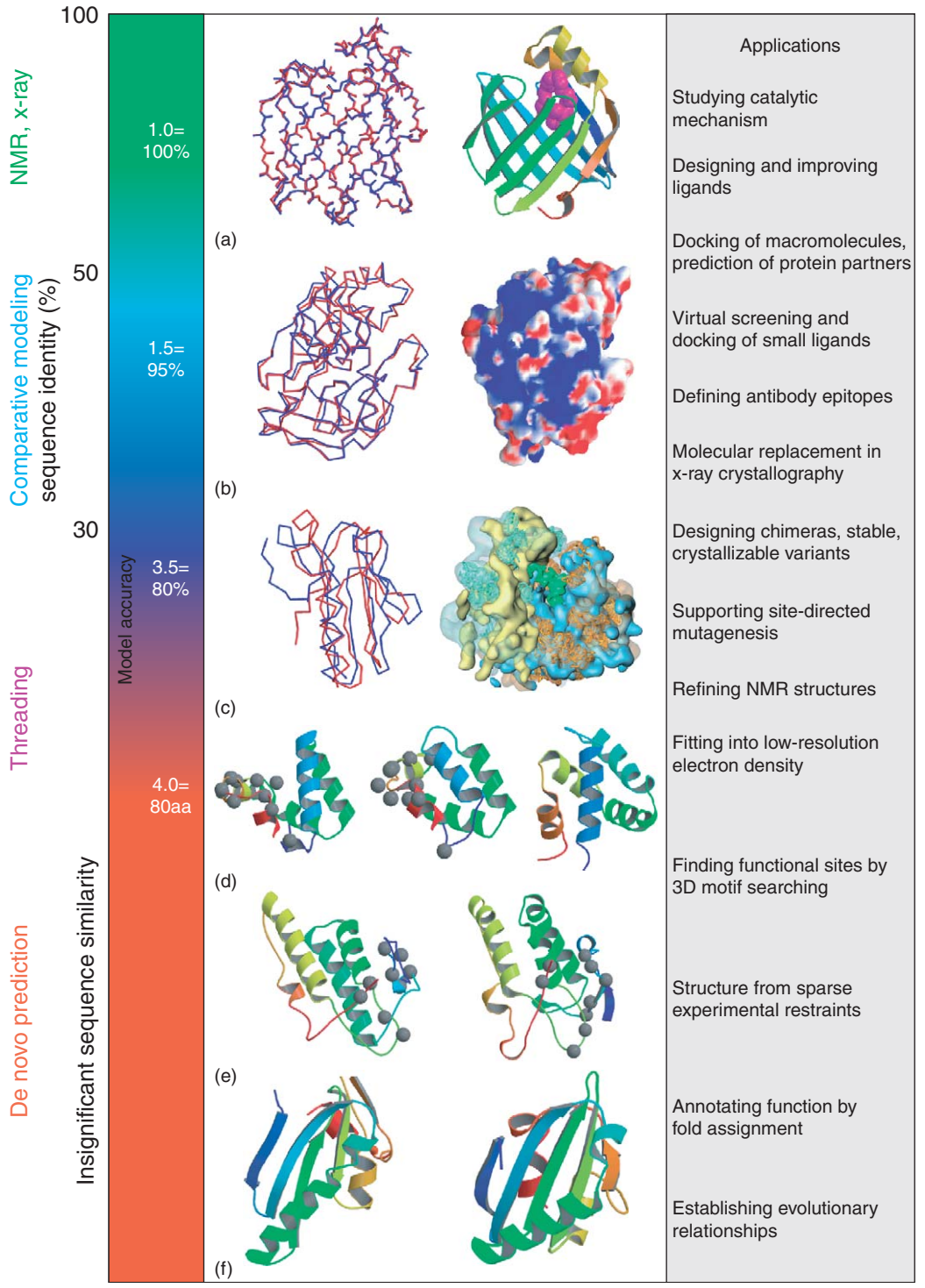

Figure 4 Accuracy and applications of protein structure models.[9] Shown are the different ranges of applicability of comparative protein structure modeling, threading, and de novo structure prediction, their corresponding accuracies, and their sample applications.

the functional regions, such as active sites, tend to be more conserved in evolution than the rest of the fold. Even low-accuracy comparative models may be useful, for example, for assigning the fold of a protein. Fold assignment can be very helpful in drug discovery, because it can shortcut the search for leads by pointing to compounds that have been previously developed for other members of the same family.[183,184]

4.10.8.1 Comparative Models versus Experimental Structures in Virtual Screening

The remainder of this review focuses on the use of comparative models for ligand docking (*see* also Chapter 4.19.2.5).[185–187] It is widely accepted that docking to comparative models is more challenging and less successful than docking to crystallographic structures. However, it seems that surprisingly little work has been done to obtain quantitative information about the accuracy of docking to comparative models, to determine in detail why the results are inferior to those obtained with crystal structures, and to improve methods for docking to comparative models.

We begin our discussion with a study by McGovern and Shoichet[188] that compared the success of docking against three different conformations of 10 enzymes: holo (ligand-bound), apo, and homology modeled. All 10 enzymes had known structures in both the holo and apo form. Comparative models for each of these enzymes were taken from MODBASE, a database of comparative models for all protein sequences that are detectably related to at least one known structure. The models were based on single template structures with sequence identities in the range of 28–87%. Each enzyme had multiple known inhibitors in the MDL Drug Data Report (MDDR) database, a library of drug-like molecules where each molecule has been annotated by the receptor to which it binds. Success of the docking, carried out with the Shoichet group's version of DOCK,[189,190] was assessed by enrichment: the ability to distinguish known inhibitors from a large set of ~100 000 'decoys' relative to random selection. As might be expected, the holo structures were the best at selecting the known ligands from among the MDDR decoys based on the docking score. Unexpectedly, the comparative models often ranked known ligands among the top-scoring database molecules; in four targets, the enrichment was 20 times higher than expected by chance.[188] In one case, purine nucleoside phosphorylase, the modeled structure actually performed better than the holo structure. For the comparative model, 25% of the known ligands were found in the top 1.2% of the ranked database, whereas for the holo conformation, 2.8% of the ranked list had to be searched before 25% of the ligands were found. In another example, the holo structure of thymidylate synthase correctly recognized ligands similar in size to the ligand captured in the x-ray structure, but not ligands that were markedly different from it. In contrast, the binding sites in the modeled conformations were more spacious and could in fact correctly detect and accommodate larger ligands than the holo receptor (**Figure 5**). Thus, it appears that, while x-ray crystallographic structures remain the first choice in docking, many comparative models seem sufficiently accurate to rank highly known ligands from among a very large list of possible alternatives.

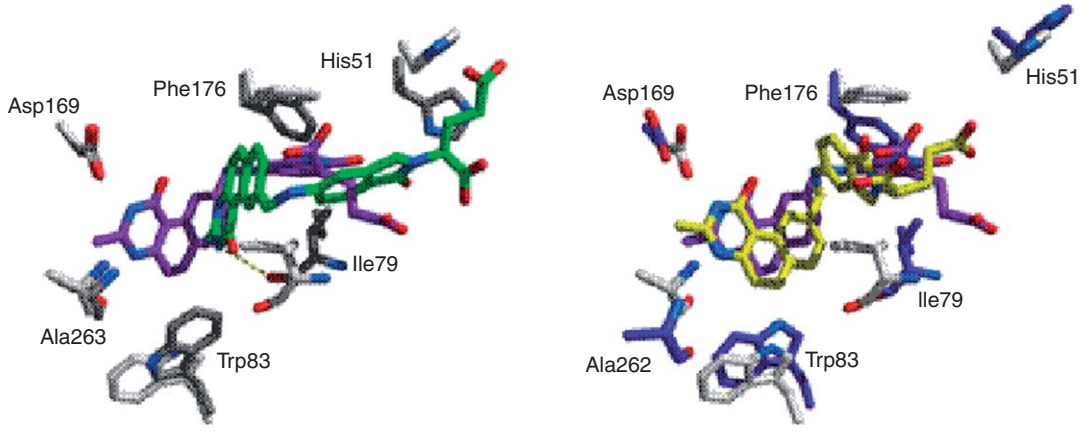

Figure 5 Docking predictions for thymidylate synthase. Shown are the x-ray structure of the holo receptor in gray, the modeled receptor in blue, the docked conformation of the ligand in the holo structure in green, and the docked conformation of the ligand in the modeled structure in yellow. A second holo complex, not used for docking but bound to a larger ligand, is also shown with protein atoms in white and ligand atoms in purple. The ligand in the holo receptor was smaller in size than many of the known ligands in the database. Consequently, while the holo structure yielded better enrichment of ligands that were similar to the native ligand, it was unable to dock larger ligands correctly. The modeled receptors, in contrast, with their more spacious binding sites, showed better competence in such cases. (Courtesy of Brian Shoichet.)

4.10.8.2 Use of Comparative Models to Obtain Novel Drug Leads

Despite problems with comparative modeling and ligand docking, comparative models have been successfully used in practice in conjunction with virtual screening to identify novel inhibitors. We briefly review a few of these success stories' to highlight the potential of the combined comparative modeling and ligand-docking approach to drug discovery (*see* 4.19 Virtual Screening).

Comparative models have been employed to aid rational drug design against parasites for more than 20 years.[122,191–193] As early as 1993, Ring *et al.*[122] used comparative models for computational docking studies that identified low micromolar nonpeptidic inhibitors of proteases in malarial and schistosome parasite lifecycles. Li *et al.*[191] subsequently used similar methods to develop nanomolar inhibitors of falcipain that are active against chloroquine-resistant strains of malaria. In a study by Selzer *et al.*[193] comparative models were used to predict new nonpeptide inhibitors of cathepsin L-like cysteine proteases in *Leishmania major*. Sixty-nine compounds were selected by DOCK 3.5 as strong binders to a comparative model of protein cpB, and of these, 21 had experimental IC_{50} values below 100 mmol L^{-1}. Finally, in a recent study by Que *et al.*[192] comparative models were used to rationalize ligand-binding affinities of cysteine proteases in *Entamoeba histolytica*. Specifically, this work provided an explanation for why proteases ACP1 and ACP2 had substrate specificity similar to that of cathepsin B, although their overall structure is more similar to that of cathepsin D.

Enyedy *et al.*[194] discovered 15 new inhibitors of matriptase by docking against its comparative model. The comparative model employed thrombin as the template, sharing only 34% sequence identity with the target sequence. Moreover, some residues in the binding site are significantly different; a trio of charged Asp residues in matriptase correspond to 1 Tyr and 2 Trp residues in thrombin. Thrombin was chosen as the template, in part because it prefers substrates with positively charged residues at the P1 position, as does matriptase. The comparative model was constructed using MODELLER and refined with MD simulations in CHARMM. The National Cancer Institute database was used for virtual screening that targeted the S1 site with the DOCK program. The 2000 best-scoring compounds were manually inspected to identify positively charged ligands (the S1 site is negatively charged), and 69 compounds were experimentally screened for inhibition, identifying the 15 inhibitors. One of them, hexamidine, was used as a lead to identify additional compounds selective for matriptase relative to thrombin. The Wang group has also used similar methods to discover seven new, low-micromolar inhibitors of Bcl-2, using a comparative model based on the NMR solution structure of Bcl-X_L.[195]

Schapira *et al.*[196] discovered a novel inhibitor of a retinoic acid receptor by virtual screening using a comparative model. In this case, the target (RAR-α) and template (RAR-γ) are very closely related; only three residues in the binding site are not conserved. The ICM program was used for virtual screening of ligands from the Available Chemicals Directory (ACD). The 5364 high-scoring compounds identified in the first round were subsequently docked into a full atom representation of the receptor with flexible side chains to obtain a final set of 300 good-scoring hits. These compounds were then manually inspected to choose the final 30 for testing. Two novel agonists were identified, with 50-nanomolar activity.

Zuccotto *et al.*[197] identified novel inhibitors of dihydrofolate reductase (DHFR) in *Trypanosoma cruzi* (the parasite that causes Chagas disease) by docking into a comparative model based on ~50% sequence identity to DHFR in *L. major*, a related parasite. The virtual screening procedure used DOCK for rigid docking of over 50 000 selected compounds from the Cambridge Structural Database (CSD). Visual inspection of the top 100 hits was used to select 36 compounds for experimental testing. This work identified several novel scaffolds with micromolar IC_{50} values. The authors report attempting to use virtual screening results to identify compounds with greater affinity for *T. cruzi* DHFR than human DHFR, but it is not clear how successful they were.

Following the recent outbreak of the severe acute respiratory syndrome (SARS) in 2003, Anand *et al.*[198] used the experimentally determined structures of the main protease from human coronavirus (M^{PRO}) and an inhibitor complex of porcine coronavirus (transmissible gastroenteritis virus, TGEV) M^{PRO} to calculate a comparative model of the SARS coronavirus M^{PRO}. This model then provided a basis for the design of anti-SARS drugs. In particular, a comparison of the active site residues in these and other related structures suggested that the AG7088 inhibitor of the human rhinovirus type 2 3C protease is a good starting point for design of anticoronaviral drugs.[199]

Comparative models of protein kinases combined with virtual screening have also been intensely used for drug discovery.[200–204] The >500 kinases in the human genome, the relatively small number of experimental structures available, and the high level of conservation around the important adenosine triphosphate-binding site make comparative modeling an attractive approach toward structure-based drug discovery.

G protein-coupled receptors are another interesting class of proteins that in principle allow drug discovery through comparative modeling.[205–209] Approximately 40% of current drug targets belong to this class of proteins. However,

these proteins have been extremely difficult to crystallize and most comparative modeling has been based on the atomic resolution structure of the bovine rhodopsin.[210] Despite this limitation, a rather extensive test of docking methods with rhodopsin-based comparative models shows encouraging results (*see* 4.26 Seven Transmembrane G Protein-Coupled Receptors: Insights for Drug Design from Structure and Modeling).

4.10.9 Future Directions

Although reports of successful virtual screening against comparative models are encouraging, such efforts are not yet a routine part of rational drug design. Even the successful efforts appear to rely strongly on visual inspection of the docking results. Much work remains to be done to improve the accuracy, efficiency, and robustness of docking against comparative models. Despite assessments of relative successes of docking against comparative models and native x-ray structures,[188,202] surprisingly little has been done to compare the accuracy achievable by different approaches to comparative modeling and to identify the specific structural reasons why comparative models generally produce less accurate virtual screening results than the holo structures. Among the many issues that deserve consideration are the following:

- The inclusion of cofactors and bound water molecules in protein receptors is often critical for success of virtual screening; however, cofactors are not routinely included in comparative models
- Most docking programs currently retain the protein receptor in a completely rigid conformation. While this approach is appropriate for 'lock-and-key' binding modes, it does not work when the ligand induces conformational changes in the receptor upon binding. A flexible receptor approach is necessary to address such induced-fit cases[211,212]
- The accuracy of comparative models is frequently judged by the C^{α} root mean square error or other similar measures of backbone accuracy. For virtual screening, however, the precise positioning of side chains in the binding site is likely to be critical; measures of accuracy for binding sites are needed to help evaluate the suitability of comparative modeling algorithms for constructing models for docking
- Knowledge of known inhibitors, either for the target protein or the template, should help to evaluate and improve virtual screening against comparative models. For example, comparative models constructed from holo' template structures implicitly preserve some information about the ligand-bound receptor conformation
- Improvement in the accuracy of models produced by comparative modeling will require methods that finely sample protein conformational space using a free energy or scoring function that has sufficient accuracy to distinguish the native structure from the nonnative conformations. Despite many years of development of molecular simulation methods, attempts to refine models that are already relatively close to the native structure have met with relatively little success. This failure is likely to be due in part to inaccuracies in the scoring functions used in the simulations, particularly in the treatment of electrostatics and solvation effects. A combination of physics-based energy function with the statistical information extracted from known protein structures may provide a route to the development of improved scoring functions
- Improvements in sampling strategies are also likely to be necessary, for both comparative modeling and flexible docking

4.10.10 Automation and Availability of Resources for Comparative Modeling and Ligand Docking

Given the increasing number of target sequences for which no experimentally determined structures are available, drug discovery stands to gain immensely from comparative modeling and other in silico methods. Despite unsolved problems in virtually every step of comparative modeling and ligand docking, it is highly desirable to automate the whole process, starting with the target sequence and ending with a ranked list of its putative ligands. Automation encourages development of better methods, improves their testing, allows application on a large scale, and makes the technology more accessible to both experts and nonspecialists alike. Through large-scale application, new questions, such as those about ligand-binding specificity, can in principle be addressed. Enabling a wider community to use the methods provides useful feedback and resources toward the development of the next generation of methods.

There are a number of servers for automated comparative modeling (Table 1). However, in spite of automation, the process of calculating a model for a given sequence, refining its structure, as well as visualizing and analyzing its family members in the sequence and structure spaces can involve the use of scripts, local programs, and servers scattered across the internet and not necessarily interconnected. In addition, manual intervention is generally still needed to

Table 1 Programs and web servers useful in comparative protein structure modeling

Name	*World Wide Web address*
Databases	
BAliBASE[222]	http://bips.u-strasbg.fr/en/Products/Databases/BAliBASE/
CATH[20]	http://www.biochem.ucl.ac.uk/bsm/cath/
DBALI[215]	http://www.salilab.org/dbali
GENBANK[14]	http://www.ncbi.nlm.nih.gov/Genbank/
GENECENSUS[223]	http://bioinfo.mbb.yale.edu/genome/
MODBASE[32]	http://www.salilab.org/modbase/
PDB[12]	http://www.rcsb.org/pdb/
PFAM[17]	http://www.sanger.ac.uk/Software/Pfam/
SCOP[19]	http://scop.mrc-lmb.cam.ac.uk/scop/
SwissProt[224]	http://www.expasy.org
Uniprot[13]	http://www.uniprot.org
Template search	
123D[225]	http://123d.ncifcrf.gov/
3D pssm[71]	http://www.sbg.bio.ic.ac.uk/~3dpssm
BLAST[36]	http://www.ncbi.nlm.nih.gov/BLAST/
DALI[33]	http://www2.ebi.ac.uk/dali/
FastA[226]	http://www.ebi.ac.uk/fasta33/
FFAS03[60]	http://ffas.ljcrf.edu/
PREDICTPROTEIN[227]	http://cubic.bioc.columbia.edu/predictprotein/
PROSPECTOR[67]	http://www.bioinformatics.buffalo.edu/current_buffalo/skolnick/prospector.html
PSIPRED[228]	http://bioinf.cs.ucl.ac.uk/psipred/
RAPTOR[68]	http://genome.math.uwaterloo.ca/~raptor/
SUPERFAMILY[229]	http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/
SAM-T02[69]	http://www.soe.ucsc.edu/research/compbio/HMM-apps/
SP3[53]	http://phyyz4.med.buffalo.edu/
SPARKS2[230]	http://phyyz4.med.buffalo.edu/
THREADER[231]	http://bioinf.cs.ucl.ac.uk/threader/threader.html
UCLA-DOE FoLD SERVER[232]	http://fold.doe-mbi.ucla.edu
Target–template alignment	
BCM SERVERF[233]	http://searchlauncher.bcm.tmc.edu
BLOCK MAKERF[234]	http://blocks.fhcrc.org/
CLUSTALW[235]	http://www2.ebi.ac.uk/clustalw/
COMPASS[57]	ftp://iole.swmed.edu/pub/compass/
FUGUE[63]	http://www-cryst.bioc.cam.ac.uk/fugue
MULTALIN[236]	http://prodes.toulouse.inra.fr/multalin/
MUSCLE[237]	http://www.drive5.com/muscle
SALIGN[213]	http://www.salilab.org/modeller
SEA[238]	http://ffas.ljcrf.edu/sea/
TCOFFEE[239]	http://www.ch.embnet.org/software/TCoffee.html
USC SEQALN[240]	http://www-hto.usc.edu/software/seqaln
Modeling	
3d-jigsaw[86]	http://www.bmm.icnet.uk/servers/3djigsaw/
COMPOSER[83]	http://www.tripos.com
CONGEN[121]	http://www.congenomics.com/
ICM[123]	http://www.molsoft.com
JACKAL[241]	http://trantor.bioc.columbia.edu/programs/jackal/
DISCOVERY STUDIO	http://www.accelrys.com
MODELLER[96]	http://www.salilab.org/modeller/

continued

Table 1 Continued

Name	*World Wide Web address*
SYBYL	http://www.tripos.com
SCWRL[147]	http://dunbrack.fccc.edu/SCWRL3.php
SNPWEB[213]	http://salilab.org/snpweb
SWISS-MODEL[87]	http://www.expasy.org/swissmod
WHAT IF[242]	http://www.cmbi.kun.nl/whatif/
Prediction of model errors	
ANOLEA[164]	http://protein.bio.puc.cl/cardex/servers/
AQUA[243]	http://urchin.bmrb.wisc.edu/~jurgen/aqua/
BIOTECH[244]	http://biotech.embl-heidelberg.de:8400
ERRAT[245]	http://www.doe-mbi.ucla.edu/Services/ERRAT/
PROCHECK[167]	http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
ProsaII[162]	http://www.came.sbg.ac.at
PROVE[246]	http://www.ucmb.ulb.ac.be/UCMB/PROVE
SQUID[247]	http://www.ysbl.york.ac.uk/~oldfield/squid/
VERIFY3D[160]	http://www.doe-mbi.ucla.edu/Services/Verify_3D/
WHATCHECK[168]	http://www.cmbi.kun.nl/gv/whatcheck/
Methods evaluation	
CAFASP[171]	http://cafasp.bioinfo.pl
CASP[248]	http://predictioncenter.llnl.gov
CASA[249]	http://capb.dbi.udel.edu/casa
EVA[174]	http://cubic.bioc.columbia.edu/eva/
LiveBench[172]	http://bioinfo.pl/LiveBench/

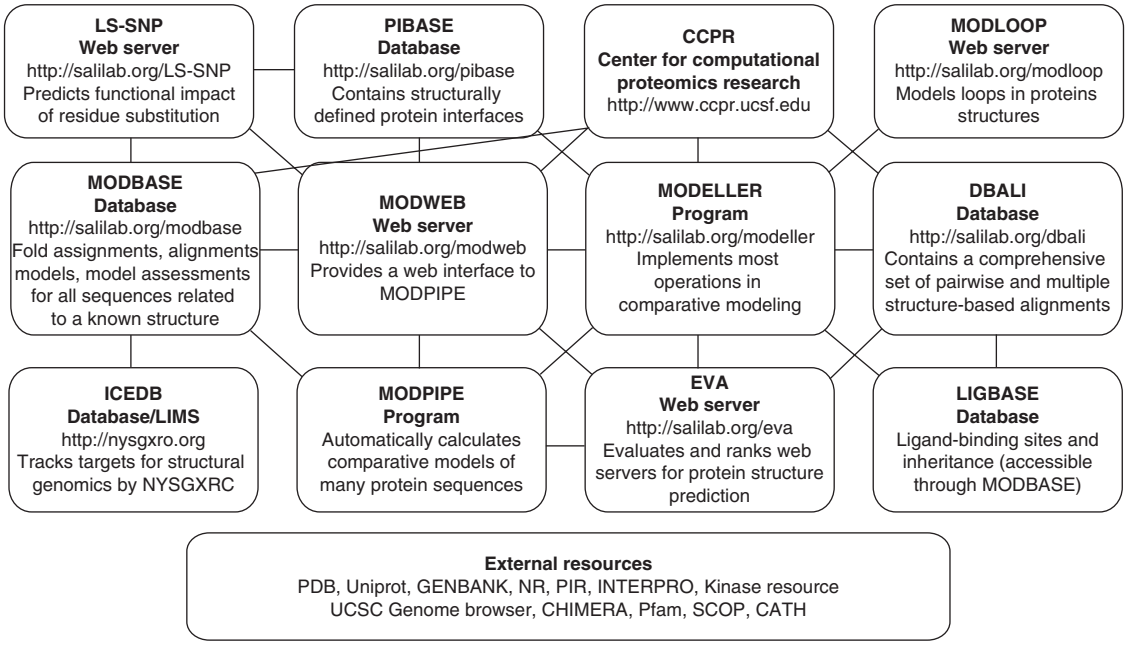

Figure 6 An integrated set of resources for comparative modeling.[32] Various databases and programs required for comparative modeling and docking are usually scattered over the internet, and require manual intervention or a good deal of expertise to be useful. Automation and integration of these resources are efficient ways to put these resources in the hands of experts and nonspecialists alike. We have outlined a comprehensive interconnected set of resources for comparative modeling and hope to integrate it with a similar effort in the area of ligand docking made by the Shoichet group.[220]

maximize the accuracy of the models in the difficult cases. The two main repositories for precomputed comparative models, SWISS-MODEL[87] and MODBASE,[31] begin to address these deficiencies. They provide access to web-based comparative modeling tools, cross-links to other sequence and structure databases, and annotations of sequences and their models.

A schematic of our own attempt at integrating several useful tools for comparative modeling is shown in Figure 6.[32,213] MODBASE is a comprehensive database that contains predicted models for domains in approximately one-half of all ~2.1 million known protein sequences. The models were calculated using MODPIPE[28,213] and MODELLER.[96] The web interface to the database allows flexible querying for fold assignments, sequence–structure alignments, models, and model assessments. An integrated sequence–structure viewer, Chimera,[214] allows inspection and analysis of the query results. Models can also be calculated using MODWEB,[213,250] a web interface to MODPIPE and stored in MODBASE to facilitate sharing, presentation, distribution, and annotation. For example, MODBASE contains binding site predictions for small ligands and a set of predicted interactions between pairs of modeled sequences from the same genome. Other resources associated with MODBASE include a comprehensive database of multiple protein structure alignments (DBALI),[215] a server for modeling of loops in protein structures (MODLOOP),[216,251] structurally defined ligand-binding sites,[217] structurally defined binary domain interfaces (PIBASE),[218,252] predictions of ligand-binding sites, interactions between yeast proteins, and functional consequences of human nsSNPs (LS-SNP).[175,219,253]

Compared to protein structure prediction, the attempts at automation and integration of resources in the field of docking for virtual screening are still in their nascent stages. One of the recent successful efforts in this direction is ZINC,[220] a publicly available database of commercially available druglike compounds. ZINC contains more than 3.3 million 'ready-to-dock' compounds organized in several subsets and allows the user to query the compounds by molecular properties and constitution. In the future, ZINC will rely on DOCKBLASTER that will enable end-users to dock the compounds against their target structures using DOCK.[189,190]

In the future, we will no doubt see efforts to improve the accuracy of comparative modeling and ligand docking. But perhaps more importantly, the two techniques will be integrated into a single protocol for more accurate and automated docking of ligands against sequences without known structures. As a result, the number and variety of applications of both comparative modeling and ligand docking will continue to increase.

Acknowledgments

This article is partially based on papers by Jacobson and Sali,[177] Fiser and Sali,[22] and Madhusudhan *et al*.[221] We also acknowledge the funds from Sandler Family Supporting Foundation, NIH R01 GM54762, P01 GM71790, P01 A135707, and U54 GM62529, as well as Sun, IBM, and Intel for hardware gifts.

References

1. Congreve, M.; Murray, C. W.; Blundell, T. L. *Drug Disc. Today* **2005**, *10*, 895–907.
2. Hardy, L., Malikayil, A. *Curr. Drug Disc.* **2003**, 15–20.
3. Lombardino, J. G.; Lowe, J. A., III *Nat. Rev. Drug Disc.* **2004**, *3*, 853–862.
4. van Dongen, M.; Weigelt, J.; Uppenberg, J.; Schultz, J.; Wikstrom, M. *Drug Disc. Today* **2002**, 7, 471–478.
5. Maryanoff, B. E. *J. Med. Chem.* **2004**, *47*, 769–787.
6. Pollack, V. A.; Savage, D. M.; Baker, D. A.; Tsaparikos, K. E.; Sloan, D. E.; Moyer, J. D.; Barbacci, E. G.; Pustilnik, L. R.; Smolarek, T. A.; Davis, J. A. et al. *J. Pharmacol. Exp. Ther.* **1999**, *291*, 739–748.
7. von Itzstein, M.; Wu, W. Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Van Phan, T.; Smythe, M. L.; White, H. F.; Oliver, S. W. et al. *Nature* **1993**, *363*, 418–423.
8. Zimmermann, J.; Caravatti, G.; Mett, H.; Meyer, T.; Muller, M.; Lydon, N. B.; Fabbro, D. *Arch. Pharm. (Weinheim)* **1996**, *329*, 371–376.
9. Baker, D.; Sali, A. *Science* **2001**, *294*, 93–96.
10. Arzt, S.; Beteva, A.; Cipriani, F.; Delageniere, S.; Felisaz, F.; Forstner, G.; Gordon, E.; Launer, L.; Lavault, B.; Leonard, G. et al. *Prog. Biophys. Mol. Biol.* **2005**, *89*, 124–152.
11. Pusey, M. L.; Liu, Z. J.; Tempel, W.; Praissman, J.; Lin, D.; Wang, B. C.; Gavira, J. A.; Ng, J. D. *Prog. Biophys. Mol. Biol.* **2005**, *88*, 359–386.
12. Deshpande, N.; Addess, K. J.; Bluhm, W. F.; Merino-Ott, J. C.; Townsend-Merino, W.; Zhang, Q.; Knezevich, C.; Xie, L.; Chen, L.; Feng, Z. et al. *Nucleic Acids Res.* **2005**, *33*, D233–D237.
13. Bairoch, A.; Apweiler, R.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M. et al. *Nucleic Acids Res.* **2005**, *33*, D154–D159.
14. Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Wheeler, D. L. *Nucleic Acids Res.* **2005**, *33*, D34–D38.
15. Chandonia, J. M.; Brenner, S. E. *Proteins* **2005**, *58*, 166–179.
16. Vitkup, D.; Melamud, E.; Moult, J.; Sander, C. *Nat. Struct. Biol.* **2001**, *8*, 559–566.
17. Bateman, A.; Coin, L.; Durbin, R.; Finn, R. D.; Hollich, V.; Griffiths-Jones, S.; Khanna, A.; Marshall, M.; Moxon, S.; Sonnhammer, E. L. et al. *Nucleic Acids Res.* **2004**, *32*, D138–D141.

18. Mulder, N. J.; Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Binns, D.; Bradley, P.; Bork, P.; Bucher, P.; Cerutti, L. et al. *Nucleic Acids Res.* **2005**, *33*, D201–D205.
19. Andreeva, A.; Howorth, D.; Brenner, S. E.; Hubbard, T. J.; Chothia, C.; Murzin, A. G. *Nucleic Acids Res.* **2004**, *32*, D226–D229.
20. Pearl, F.; Todd, A.; Sillitoe, I.; Dibley, M.; Redfern, O.; Lewis, T.; Bennett, C.; Marsden, R.; Grant, A.; Lee, D. et al. *Nucleic Acids Res.* **2005**, *33*, D247–D251.
21. Godzik, A. *Methods Biochem. Anal.* **2003**, *44*, 525–546.
22. Fiser, A.; Sali, A. *Methods Enzymol.* **2003**, *374*, 461–491.
23. Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291–325.
24. Hillisch, A.; Pineda, L. F.; Hilgenfeld, R. *Drug Disc. Today* **2004**, *9*, 659–669.
25. Jorgensen, W. L. *Science* **2004**, *303*, 1813–1818.
26. Bradley, P.; Misura, K. M.; Baker, D. *Science* **2005**, *309*, 1868–1871.
27. Chothia, C.; Lesk, A. M. *EMBO J.* **1986**, *5*, 823–826.
28. Sanchez, R.; Sali, A. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 13597–13602.
29. Sanchez, R.; Pieper, U.; Melo, F.; Eswar, N.; Marti-Renom, M. A.; Madhusudhan, M. S.; Mirkovic, N.; Sali, A. *Nat. Struct. Biol.* **2000**, 7, 986–990.
30. Sali, A. *Nat. Struct. Biol.* **1998**, *5*, 1029–1032.
31. Sali, A. *Nat. Struct. Biol.* **2001**, *8*, 482–484.
32. Pieper, U.; Eswar, N.; Braberg, H.; Madhusudhan, M. S.; Davis, F. P.; Stuart, A. C.; Mirkovic, N.; Rossi, A.; Marti-Renom, M. A.; Fiser, A. et al. *Nucleic Acids Res.* **2004**, *32*, D217–D222.
33. Dietmann, S.; Park, J.; Notredame, C.; Heger, A.; Lappe, M.; Holm, L. *Nucleic Acids Res.* **2001**, *29*, 55–57.
34. Rost, B. *Protein Eng.* **1999**, *12*, 85–94.
35. Pearson, W. R. *Methods Mol. Biol.* **1994**, *24*, 307–331.
36. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. *J. Nucleic Acids Res.* **1997**, *25*, 3389–3402.
37. Brenner, S. E.; Chothia, C.; Hubbard, T. J. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 6073–6078.
38. Saqi, M. A.; Russell, R. B.; Sternberg, M. J. *Protein Eng.* **1998**, *11*, 627–630.
39. Gribskov, M.; McLachlan, A. D.; Eisenberg, D. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 4355–4358.
40. Henikoff, J. G.; Henikoff, S. *Comput. Appl. Biosci.* **1996**, *12*, 135–243.
41. Henikoff, S.; Henikoff, J. G. *J. Mol. Biol.* **1994**, *243*, 574–578.
42. Eddy, S. R. *Bioinformatics* **1998**, *14*, 755–763.
43. Krogh, A.; Brown, M.; Mian, I. S.; Sjolander, K.; Haussler, D. *J. Mol. Biol.* **1994**, *235*, 1501–1531.
44. Lindahl, E.; Elofsson, A. *J. Mol. Biol.* **2000**, *295*, 613–625.
45. Park, J.; Karplus, K.; Barrett, C.; Hughey, R.; Haussler, D.; Hubbard, T.; Chothia, C. *J. Mol. Biol.* **1998**, *284*, 1201–1210.
46. Sauder, J. M.; Arthur, J. W.; Dunbrack, R. L., Jr. *Proteins* **2000**, *40*, 6–22.
47. Marti-Renom, M. A.; Madhusudhan, M. S.; Sali, A. *Protein Sci.* **2004**, *13*, 1071–1087.
48. Karplus, K.; Barrett, C.; Hughey, R. *Bioinformatics* **1998**, *14*, 846–856.
49. Eswar, N., Madhusudhan, M. S., Marti-Renom, M. A., Sali, A.; 8v1 ed. 2005. http://salilab.org/modeller (accessed Aug 2006).
50. Edgar, R. C.; Sjolander, K. *Bioinformatics* **2004**, *20*, 1301–1308.
51. Ohlson, T.; Wallner, B.; Elofsson, A. *Proteins* **2004**, *57*, 188–197.
52. Wang, G.; Dunbrack, R. L., Jr. *Protein Sci.* **2004**, *13*, 1612–1626.
53. Zhou, H.; Zhou, Y. *Proteins* **2005**, *58*, 321–328.
54. Panchenko, A. R. *Nucleic Acids Res.* **2003**, *31*, 683–689.
55. Pietrokovski, S. *Nucleic Acids Res.* **1996**, *24*, 3836–3845.
56. Rychlewski, L.; Zhang, B.; Godzik, A. *Fold Des.* **1998**, *3*, 229–238.
57. Sadreyev, R.; Grishin, N. *J. Mol. Biol.* **2003**, *326*, 317–336.
58. von Ohsen, N.; Sommer, I.; Zimmer, R. *Pac. Symp. Biocomput.* **2003**, 252–263.
59. Yona, G.; Levitt, M. *J. Mol. Biol.* **2002**, *315*, 1257–1275.
60. Jaroszewski, L.; Rychlewski, L.; Li, Z.; Li, W.; Godzik, A. *Nucleic Acids Res.* **2005**, *33*, W284–W288.
61. McGuffin, L. J.; Jones, D. T. *Bioinformatics* **2003**, *19*, 874–981.
62. Karchin, R.; Cline, M.; Mandel-Gutfreund, Y.; Karplus, K. *Proteins* **2003**, *51*, 504–514.
63. Shi, J.; Blundell, T. L.; Mizuguchi, K. *J. Mol. Biol.* **2001**, *310*, 243–257.
64. Bowie, J. U.; Luthy, R.; Eisenberg, D. *Science* **1991**, *253*, 164–170.
65. Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859–883.
66. Sippl, M. J. *Curr. Opin. Struct. Biol.* **1995**, *5*, 229–235.
67. Skolnick, J.; Kihara, D. *Proteins* **2001**, *42*, 319–331.
68. Xu, J.; Li, M.; Kim, D.; Xu, Y. *J. Bioinform. Comput. Biol.* **2003**, *1*, 95–117.
69. Karplus, K.; Karchin, R.; Draper, J.; Casper, J.; Mandel-Gutfreund, Y.; Diekhans, M.; Hughey, R. *Proteins* **2003**, *53*, 491–496.
70. Jones, D. T. *J. Mol. Biol.* **1999**, *287*, 797–815.
71. Kelley, L. A.; MacCallum, R. M.; Sternberg, M. J. *J. Mol. Biol.* **2000**, *299*, 499–520.
72. John, B.; Sali, A. *Nucleic Acids Res.* **2003**, *31*, 3982–3992.
73. Moult, J. *Curr. Opin. Struct. Biol.* **2005**, *15*, 285–289.
74. Sanchez, R.; Sali, A. *Curr. Opin. Struct. Biol.* **1997**, 7, 206–214.
75. Sanchez, R.; Sali, A. *Proteins* **1997**, 50–58.
76. Srinivasan, N.; Blundell, T. L. *Protein Eng.* **1993**, *6*, 501–512.
77. Bajorath, J.; Aruffo, A. *Bioconjug. Chem.* **1994**, *5*, 173–181.
78. Blundell, T. L.; Sibanda, B. L.; Sternberg, M. J.; Thornton, J. M. *Nature* **1987**, *326*, 347–352.
79. Browne, W. J.; North, A. C.; Phillips, D. C.; Brew, K.; Vanaman, T. C.; Hill, R. L. *J. Mol. Biol.* **1969**, *42*, 65–86.
80. Johnson, M. S.; Srinivasan, N.; Sowdhamini, R.; Blundell, T. L. *Crit. Rev. Biochem. Mol. Biol.* **1994**, *29*, 1–68.
81. Greer, J. *J. Mol. Biol.* **1981**, *153*, 1027–1042.
82. Nagarajaram, H. A.; Reddy, B. V.; Blundell, T. L. *Protein Eng.* **1999**, *12*, 1055–1062.
83. Sutcliffe, M. J.; Haneef, I.; Carney, D.; Blundell, T. L. *Protein Eng.* **1987**, *1*, 377–384.

84. Sutcliffe, M. J.; Hayes, F. R.; Blundell, T. L. *Protein Eng.* **1987**, *1*, 385–392.
85. Topham, C. M.; McLeod, A.; Eisenmenger, F.; Overington, J. P.; Johnson, M. S.; Blundell, T. L. *J. Mol. Biol.* **1993**, *229*, 194–220.
86. Bates, P. A.; Kelley, L. A.; MacCallum, R. M.; Sternberg, M. J. *Proteins* **2001**, 39–46.
87. Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. *Nucleic Acids Res.* **2003**, *31*, 3381–3385.
88. Bystroff, C.; Baker, D. *J. Mol. Biol.* **1998**, *281*, 565–577.
89. Claessens, M.; Van Cutsem, E.; Lasters, I.; Wodak, S. *Protein Eng.* **1989**, *2*, 335–345.
90. Jones, T. A.; Thirup, S. *EMBO J.* **1986**, *5*, 819–822.
91. Levitt, M. *J. Mol. Biol.* **1992**, *226*, 507–533.
92. Unger, R.; Harel, D.; Wherland, S.; Sussman, J. L. *Proteins* **1989**, *5*, 355–373.
93. Aszodi, A.; Taylor, W. R. *Fold Des.* **1996**, *1*, 325–334.
94. Brocklehurst, S. M.; Perham, R. N. *Protein Sci.* **1993**, *2*, 626–639.
95. Havel, T. F.; Snow, M. E. *J. Mol. Biol.* **1991**, *217*, 1–7.
96. Sali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779–815.
97. Srinivasan, S.; March, C. J.; Sudarsanam, S. *Protein Sci.* **1993**, *2*, 277–289.
98. Fiser, A.; Do, R. K.; Sali, A. *Protein Sci.* **2000**, *9*, 1753–1773.
99. Fiser, A.; Feig, M.; Brooks, C. L., III; Sali, A. *Acc. Chem. Res.* **2002**, *35*, 413–421.
100. MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S. et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
101. Sali, A.; Overington, J. P. *Protein Sci.* **1994**, *3*, 1582–1596.
102. Marti-Renom, M. A.; Madhusudhan, M. S.; Fiser, A.; Rost, B.; Sali, A. *Struct. (Camb.)* **2002**, *10*, 435–440.
103. Wallner, B.; Elofsson, A. *Protein Sci.* **2005**, *14*, 1315–1327.
104. Kabsch, W.; Sander, C. *Proc. Natl. Acad. Sci. USA* **1984**, *81*, 1075–1078.
105. Mezei, M. *Protein Eng.* **1998**, *11*, 411–414.
106. Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A. *Proteins* **2004**, *55*, 351–367.
107. Chothia, C.; Lesk, A. M. *J. Mol. Biol.* **1987**, *196*, 901–917.
108. Moult, J.; James, M. N. *Proteins* **1986**, *1*, 146–163.
109. Bruccoleri, R. E.; Karplus, M. *Biopolymers* **1987**, *26*, 137–168.
110. Shenkin, P. S.; Yarmush, D. L.; Fine, R. M.; Wang, H. J.; Levinthal, C. *Biopolymers* **1987**, *26*, 2053–2085.
111. van Vlijmen, H. W.; Karplus, M. *J. Mol. Biol.* **1997**, *267*, 975–1001.
112. Deane, C. M.; Blundell, T. L. *Protein Sci.* **2001**, *10*, 599–612.
113. Sibanda, B. L.; Blundell, T. L.; Thornton, J. M. *J. Mol. Biol.* **1989**, *206*, 759–777.
114. Chothia, C.; Lesk, A. M.; Tramontano, A.; Levitt, M.; Smith-Gill, S. J.; Air, G.; Sheriff, S.; Padlan, E. A.; Davies, D.; Tulip, W. R. et al. *Nature* **1989**, *342*, 877–883.
115. Rufino, S. D.; Donate, L. E.; Canard, L. H.; Blundell, T. L. *J. Mol. Biol.* **1997**, *267*, 352–367.
116. Oliva, B.; Bates, P. A.; Querol, E.; Aviles, F. X.; Sternberg, M. J. *J. Mol. Biol.* **1997**, *266*, 814–830.
117. Ring, C. S.; Kneller, D. G.; Langridge, R.; Cohen, F. E. *J. Mol. Biol.* **1992**, *224*, 685–699.
118. Fidelis, K.; Stern, P. S.; Bacon, D.; Moult, J. *Protein Eng.* **1994**, 7, 953–960.
119. Lessel, U.; Schomburg, D. *Protein Eng.* **1994**, 7, 1175–1187.
120. Fine, R. M.; Wang, H.; Shenkin, P. S.; Yarmush, D. L.; Levinthal, C. *Proteins* **1986**, *1*, 342–362.
121. Bruccoleri, R. E.; Karplus, M. *Biopolymers* **1990**, *29*, 1847–1862.
122. Ring, C. S.; Sun, E.; McKerrow, J. H.; Lee, G. K.; Rosenthal, P. J.; Kuntz, I. D.; Cohen, F. E. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 3583–3587.
123. Abagyan, R.; Totrov, M. *J. Mol. Biol.* **1994**, *235*, 983–1002.
124. Collura, V.; Higo, J.; Garnier, J. *Protein Sci.* **1993**, *2*, 1502–1510.
125. Higo, J.; Collura, V.; Garnier, J. *Biopolymers* **1992**, *32*, 33–43.
126. Zheng, Q.; Rosenfeld, R.; Vajda, S.; DeLisi, C. *Protein Sci.* **1993**, *2*, 1242–1248.
127. Koehl, P.; Delarue, M. *Nat. Struct. Biol.* **1995**, *2*, 163–170.
128. Samudrala, R.; Moult, J. *J. Mol. Biol.* **1998**, *279*, 287–302.
129. Xiang, Z.; Soto, C. S.; Honig, B. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7432–7437.
130. de Bakker, P. I.; DePristo, M. A.; Burke, D. F.; Blundell, T. L. *Proteins* **2003**, *51*, 21–40.
131. DePristo, M. A.; de Bakker, P. I.; Lovell, S. C.; Blundell, T. L. *Proteins* **2003**, *51*, 41–55.
132. Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. *Proteins-Struct. Funct. Genet.* **2002**, *48*, 404–422.
133. Dunbrack, R. L., Jr. *Curr. Opin. Struct. Biol.* **2002**, *12*, 431–440.
134. Janin, J.; Chothia, C. *Biochemistry* **1978**, *17*, 2943–2948.
135. Ponder, J. W.; Richards, F. M. *J. Mol. Biol.* **1987**, *193*, 775–791.
136. De Maeyer, M.; Desmet, J.; Lasters, I. *Fold Des.* **1997**, *2*, 53–66.
137. Dunbrack, R. L., Jr.; Cohen, F. E. *Protein Sci.* **1997**, *6*, 1661–1681.
138. Dunbrack, R. L., Jr.; Karplus, M. *J. Mol. Biol.* **1993**, *230*, 543–574.
139. Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. *Proteins* **2000**, *40*, 389–408.
140. McGregor, M. J.; Islam, S. A.; Sternberg, M. J. *J. Mol. Biol.* **1987**, *198*, 295–310.
141. Schrauber, H.; Eisenhaber, F.; Argos, P. *J. Mol. Biol.* **1993**, *230*, 592–612.
142. Tuffery, P.; Etchebest, C.; Hazout, S.; Lavery, R. *J. Biomol. Struct. Dyn.* **1991**, *8*, 1267–1289.
143. Desjarlais, J. R.; Handel, T. M. *J. Mol. Biol.* **1999**, *290*, 305–318.
144. De Filippis, V.; Sander, C.; Vriend, G. *Protein Eng.* **1994**, 7, 1203–1208.
145. Chung, S. Y.; Subbiah, S. *Pac. Symp. Biocomput.* **1996**, 126–141.
146. Cregut, D.; Liautard, J. P.; Chiche, L. *Protein Eng.* **1994**, 7, 1333–1344.
147. Canutescu, A. A.; Shelenkov, A. A.; Dunbrack, R. L., Jr. *Protein Sci.* **2003**, *12*, 2001–2014.
148. Xiang, Z.; Honig, B. *J. Mol. Biol.* **2001**, *311*, 421–430.
149. Eisenmenger, F.; Argos, P.; Abagyan, R. *J. Mol. Biol.* **1993**, *231*, 849–860.
150. Lee, G. M.; Varma, A.; Palsson, B. O. *Biotechnol. Prog.* **1991**, 7, 72–75.
151. Holm, L.; Sander, C. *Proteins* **1992**, *14*, 213–223.

152. Lasters, I.; Desmet, J. *Protein Eng.* **1993**, *6*, 717–722.
153. Looger, L. L.; Hellinga, H. W. *J. Mol. Biol.* **2001**, *307*, 429–445.
154. Hwang, J. K.; Liao, W. F. *Protein Eng.* **1995**, *8*, 363–370.
155. Koehl, P.; Delarue, M. *J. Mol. Biol.* **1994**, *239*, 249–275.
156. Bower, M. J.; Cohen, F. E.; Dunbrack, R. L., Jr. *J. Mol. Biol.* **1997**, *267*, 1268–1282.
157. Petrella, R. J.; Lazaridis, T.; Karplus, M. *Fold Des.* **1998**, *3*, 353–377.
158. Jacobson, M. P.; Kaminski, G. A.; Friesner, R. A.; Rapp, C. S. *J. Phys. Chem. B* **2002**, *106*, 11673–11680.
159. Liang, S.; Grishin, N. V. *Protein Sci.* **2002**, *11*, 322–331.
160. Luthy, R.; Bowie, J. U.; Eisenberg, D. *Nature* **1992**, *356*, 83–85.
161. Melo, F.; Sanchez, R.; Sali, A. *Protein Sci.* **2002**, *11*, 430–448.
162. Sippl, M. J. *Proteins* **1993**, *17*, 355–362.
163. Topham, C. M.; Srinivasan, N.; Thorpe, C. J.; Overington, J. P.; Kalsheker, N. A. *Protein Eng.* **1994**, 7, 869–894.
164. Melo, F.; Feytmans, E. *J. Mol. Biol.* **1998**, *277*, 1141–1152.
165. Zhou, H.; Zhou, Y. *Protein Sci.* **2002**, *11*, 2714–2726.
166. Pawlowski, K.; Bierzynski, A.; Godzik, A. *J. Mol. Biol.* **1996**, *258*, 349–366.
167. Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. *J. Appl. Crystallogr.* **1993**, *26*, 283–291.
168. Hooft, R. W.; Vriend, G.; Sander, C.; Abola, E. E. *Nature* **1996**, *381*, 272.
169. Coutsias, E. A.; Seok, C.; Jacobson, M. P.; Dill, K. A. *J. Comput. Chem.* **2004**, *25*, 510–528.
170. Zemla, A.; Venclovas, C.; Moult, J.; Fidelis, K. *Proteins* **2001**, 13–21.
171. Fischer, D.; Elofsson, A.; Rychlewski, L.; Pazos, F.; Valencia, A.; Rost, B.; Ortiz, A. R.; Dunbrack, R. L., Jr. *Proteins* **2001**, *45*, 171–183.
172. Bujnicki, J. M.; Elofsson, A.; Fischer, D.; Rychlewski, L. *Protein Sci.* **2001**, *10*, 352–361.
173. Eyrich, V. A.; Marti-Renom, M. A.; Przybylski, D.; Madhusudhan, M. S.; Fiser, A.; Pazos, F.; Valencia, A.; Sali, A.; Rost, B. *Bioinformatics* **2001**, *17*, 1242–1243.
174. Koh, I.-Y. Y.; Eyrich, V. A.; Marti-Renom, M. A.; Przybylski, D.; Madhusudhan, M. S.; Narayanan, E.; Grana, O.; Pazos, F.; Valencia, A.; Sali, A. et al. *Nucleic Acids Res.* **2003**, *31*, 3311–3315.
175. Karchin, R.; Diekhans, M.; Kelly, L.; Thomas, D. J.; Pieper, U.; Eswar, N.; Haussler, D.; Sali, A. *Bioinformatics* **2005**, *21*, 2814–2820.
176. Thiel, K. A. *Nat. Biotechnol.* **2004**, *22*, 513–519.
177. Jacobson, M. P., Sali, A. *Comparative Modeling and Its Applications to Drug Discovery*; Inpharmatica: London, 2004; Vol. 39.
178. Gao, H.; Sengupta, J.; Valle, M.; Korostelev, A.; Eswar, N.; Stagg, S. M.; Van Roey, P.; Agrawal, R. K.; Harvey, S. C.; Sali, A. et al. *Cell* **2003**, *113*, 789–801.
179. Spahn, C. M.; Beckmann, R.; Eswar, N.; Penczek, P. A.; Sali, A.; Blobel, G.; Frank, J. *Cell* **2001**, *107*, 373–386.
180. Blundell, T. L.; Johnson, M. S. *Protein Sci.* **1993**, *2*, 877–883.
181. Chakravarty, S.; Sanchez, R. *Structure (Camb.)* **2004**, *12*, 1461–1470.
182. Chakravarty, S.; Wang, L.; Sanchez, R. *Nucleic Acids Res.* **2005**, *33*, 244–259.
183. von Grotthuss, M.; Wyrwicz, L. S.; Rychlewski, L. *Cell* **2003**, *113*, 701–702.
184. Gordon, R. K.; Ginalski, K.; Rudnicki, W. R.; Rychlewski, L.; Pankaskie, M. C.; Bujnicki, J. M.; Chiang, P. K. *Eur. J. Biochem.* **2003**, *270*, 3507–3517.
185. Evers, A.; Gohlke, H.; Klebe, G. *J. Mol. Biol.* **2003**, *334*, 327–345.
186. Evers, A.; Klebe, G. *Angew. Chem. Int. Ed. Engl.* **2004**, *43*, 248–251.
187. Schafferhans, A.; Klebe, G. *J. Mol. Biol.* **2001**, *307*, 407–427.
188. McGovern, S. L.; Shoichet, B. K. *J. Med. Chem.* **2003**, *46*, 2895–2907.
189. Lorber, D. M.; Shoichet, B. K. *Protein Sci.* **1998**, 7, 938–950.
190. Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. *J. Mol. Biol.* **2002**, *322*, 339–355.
191. Li, R.; Chen, X.; Gong, B.; Selzer, P. M.; Li, Z.; Davidson, E.; Kurzban, G.; Miller, R. E.; Nuzum, E. O.; McKerrow, J. H. et al. *Bioorg. Med. Chem.* 1996, *4*, 1421–1427.
192. Que, X.; Brinen, L. S.; Perkins, P.; Herdman, S.; Hirata, K.; Torian, B. E.; Rubin, H.; McKerrow, J. H.; Reed, S. L. *Mol. Biochem. Parasitol.* **2002**, *119*, 23–32.
193. Selzer, P. M.; Chen, X.; Chan, V. J.; Cheng, M.; Kenyon, G. L.; Kuntz, I. D.; Sakanari, J. A.; Cohen, F. E.; McKerrow, J. H. *Exp. Parasitol.* **1997**, *87*, 212–221.
194. Enyedy, I. J.; Lee, S. L.; Kuo, A. H.; Dickson, R. B.; Lin, C. Y.; Wang, S. *J. Med. Chem.* **2001**, *44*, 1349–1355.
195. Enyedy, I. J.; Ling, Y.; Nacro, K.; Tomita, Y.; Wu, X.; Cao, Y.; Guo, R.; Li, B.; Zhu, X.; Huang, Y. et al. *J. Med. Chem.* **2001**, *44*, 4313–4324.
196. Schapira, M.; Raaka, B. M.; Samuels, H. H.; Abagyan, R. *BMC Struct Biol* **2001**, *1*, 1.
197. Zuccotto, F.; Zvelebil, M.; Brun, R.; Chowdhury, S. F.; Di Lucrezia, R.; Leal, I.; Maes, L.; Ruiz-Perez, L. M.; Gonzalez Pacanowska, D.; Gilbert, I. H. *Eur. J. Med. Chem.* **2001**, *36*, 395–405.
198. Anand, K.; Ziebuhr, J.; Wadhwani, P.; Mesters, J. R.; Hilgenfeld, R. *Science* **2003**, *300*, 1763–1767.
199. Rajnarayanan, R. V.; Dakshanamurthy, S.; Pattabiraman, N. *Biochem. Biophys. Res. Commun.* **2004**, *321*, 370–378.
200. Diller, D. J.; Li, R. *J. Med. Chem.* **2003**, *46*, 4638–4647.
201. Diller, D. J.; Merz, K. M., Jr. *Proteins* **2001**, *43*, 113–124.
202. Oshiro, C.; Bradley, E. K.; Eksterowicz, J.; Evensen, E.; Lamb, M. L.; Lanctot, J. K.; Putta, S.; Stanton, R.; Grootenhuis, P. D. *J. Med. Chem.* **2004**, *47*, 764–767.
203. Rockey, W. M.; Elcock, A. H. *Proteins* **2002**, *48*, 664–671.
204. Vangrevelinghe, E.; Zimmermann, K.; Schoepfer, J.; Portmann, R.; Fabbro, D.; Furet, P. *J. Med. Chem.* **2003**, *46*, 2656–2662.
205. Becker, O. M.; Shacham, S.; Marantz, Y.; Noiman, S. *Curr. Opin. Drug Disc. Dev.* **2003**, *6*, 353–361.
206. Bissantz, C.; Bernard, P.; Hibert, M.; Rognan, D. *Proteins* **2003**, *50*, 5–25.
207. Bissantz, C.; Logean, A.; Rognan, D. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1162–1176.
208. Shacham, S.; Topf, M.; Avisar, N.; Glaser, F.; Marantz, Y.; Bar-Haim, S.; Noiman, S.; Naor, Z.; Becker, O. M. *Med. Res. Rev.* **2001**, *21*, 472–483.
209. Vaidehi, N.; Floriano, W. B.; Trabanino, R.; Hall, S. E.; Freddolino, P.; Choi, E. J.; Zamanakos, G.; Goddard, W. A., III *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12622–12627.
210. Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Le Trong, I.; Teller, D. C.; Okada, T.; Stenkamp, R. E. et al. *Science* **2000**, *289*, 739–745.
211. Barril, X.; Morley, S. D. *J. Med. Chem.* **2005**, *48*, 4432–4443.

212. Carlson, H. A.; McCammon, J. A. *Mol. Pharmacol.* **2000**, *57*, 213–218.
213. Eswar, N.; John, B.; Mirkovic, N.; Fiser, A.; Ilyin, V. A.; Pieper, U.; Stuart, A. C.; Marti-Renom, M. A.; Madhusudhan, M. S.; Yerkovich, B. et al. *Nucleic Acids Res.* **2003**, *31*, 3375–3380.
214. Huang, C. C.; Novak, W. R.; Babbitt, P. C.; Jewett, A. I.; Ferrin, T. E.; Klein, T. E. *Pac. Symp. Biocomput.* **2000**, 230–241.
215. Marti-Renom, M. A.; Ilyin, V. A.; Sali, A. *Bioinformatics* **2001**, *17*, 746–747.
216. Fiser, A.; Sali, A. *Bioinformatics* **2003**, *19*, 2500–2501.
217. Stuart, A. C.; Ilyin, V. A.; Sali, A. *Bioinformatics* **2002**, *18*, 200–201.
218. Davis, F. P.; Sali, A. *Bioinformatics* **2005**, *21*, 1901–1907.
219. Mirkovic, N.; Marti-Renom, M. A.; Weber, B. L.; Sali, A.; Monteiro, A. N. *Cancer Res.* **2004**, *64*, 3790–3797.
220. Irwin, J. J.; Shoichet, B. K. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
221. Madhusudhan, M. S.; Marti-Renom, M. A.; Eswar, N.; John, B.; Pieper, U.; Karchin, R.; Shen, M. Y.; Sali, A. In *The Proteomics Protocols Handbook*; Walker, J. M., Ed.; Humana Press: Totowa, NJ, 2005, pp 831–860.
222. Thompson, J. D.; Plewniak, F.; Poch, O. *Bioinformatics* **1999**, *15*, 87–88.
223. Lin, J.; Qian, J.; Greenbaum, D.; Bertone, P.; Das, R.; Echols, N.; Senes, A.; Stenger, B.; Gerstein, M. *Nucleic Acids Res.* **2002**, *30*, 4574–4582.
224. Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I. et al. *Nucleic Acids Res.* **2003**, *31*, 365–370.
225. Alexandrov, N. N.; Nussinov, R.; Zimmer, R. M. *Pac. Symp. Biocomput.* **1996**, 53–72.
226. Pearson, W. R. *Methods Mol. Biol.* **2000**, *132*, 185–219.
227. Rost, B.; Liu, J. *Nucleic Acids Res.* **2003**, *31*, 3300–3304.
228. McGuffin, L. J.; Bryson, K.; Jones, D. T. *Bioinformatics* **2000**, *16*, 404–405.
229. Gough, J.; Karplus, K.; Hughey, R.; Chothia, C. *J. Mol. Biol.* **2001**, *313*, 903–919.
230. Zhou, H.; Zhou, Y. *Proteins* **2004**, *55*, 1005–1013.
231. Jones, D. T.; Taylor, W. R.; Thornton, J. M. *Nature* **1992**, *358*, 86–89.
232. Mallick, P.; Weiss, R.; Eisenberg, D. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 16041–16046.
233. Worley, K. C.; Culpepper, P.; Wiese, B. A.; Smith, R. F. *Bioinformatics* **1998**, *14*, 890–891.
234. Henikoff, J. G.; Pietrokovski, S.; McCallum, C. M.; Henikoff, S. *Electrophoresis* **2000**, *21*, 1700–1706.
235. Thompson, J. D.; Higgins, D. G.; Gibson, T. J. *Nucleic Acids Res.* **1994**, *22*, 4673–4680.
236. Corpet, F. *Nucleic Acids Res.* **1988**, *16*, 10881–10890.
237. Edgar, R. C. *Nucleic Acids Res.* **2004**, *32*, 1792–1797.
238. Ye, Y.; Jaroszewski, L.; Li, W.; Godzik, A. *Bioinformatics* **2003**, *19*, 742–749.
239. Notredame, C.; Higgins, D. G.; Heringa, J. *J. Mol. Biol.* **2000**, *302*, 205–217.
240. Smith, T. F.; Waterman, M. S. *J. Mol. Biol.* **1981**, *147*, 195–197.
241. Petrey, D.; Xiang, Z.; Tang, C. L.; Xie, L.; Gimpelev, M.; Mitros, T.; Soto, C. S.; Goldsmith-Fischman, S.; Kernytsky, A.; Schlessinger, A. et al. *Proteins* **2003**, *53*, 430–435.
242. Vriend, G. *J. Mol. Graph.* **1990**, *8*, 52–56.
243. Laskowski, R. A.; Rullmannn, J. A.; MacArthur, M. W.; Kaptein, R.; Thornton, J. M. *J. Biomol. NMR* **1996**, *8*, 477–486.
244. Laskowski, R. A.; MacArthur, M. W.; Thornton, J. M. *Curr. Opin. Struct. Biol.* **1998**, *8*, 631–639.
245. Colovos, C.; Yeates, T. O. *Protein Sci.* **1993**, *2*, 1511–1519.
246. Pontius, J.; Richelle, J.; Wodak, S. J. *J. Mol. Biol.* **1996**, *264*, 121–136.
247. Oldfield, T. J. *J. Mol. Graph.* **1992**, *10*, 247–252.
248. Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. *Proteins* **2003**, *53*, 334–339.
249. Kahsay, R. Y.; Wang, G.; Dongre, N.; Gao, G.; Dunbrack, R. L., Jr. *Bioinformatics* **2002**, *18*, 496–497.
250. MODWEB. http://salilab.org/modweb (accessed April 2006).
251. MODLOOP. http://salilab.org/modloop (accessed April 2006).
252. PIBASE. http://salilab.org/pibase (accessed April 2006).
253. LS-SNP. http://salilab.org/LS-SNP (accessed April 2006).

Biographies

Eswar Narayanan received his BSc degree in physics from the Loyola College, India, in 1993 for which he was awarded a Gold Medal and Scholarship for academic proficiency. After an MSc degree in Physics from the University of Hyderabad, India, he was awarded a Research Fellowship by the Indian Institute of Science, Bangalore, India, for a PhD

under the supervision of Prof C Ramakrishnan at the Molecular Biophysics Unit where he focused on the conformational analysis of protein structures. He then joined the laboratory of Prof Andrej Sali at the Rockefeller University, New York, as a Research Associate where he developed the large-scale protein structure modeling pipeline, MODPIPE. In 2003, he moved along with Prof Sali to the University of California at San Francisco, where, as an Assistant Professional Researcher, he continues to work on the development of methods for protein structure prediction and its application to modeling structures of macromolecular assemblies and modeling drug target proteins for virtual screening.

Andrej Sali received his BSc degree in chemistry from the University of Ljubljana, Slovenia, in 1987. He was awarded the Research Council of Slovenia Scholarship, the Overseas Research Students Award, and the Merck Sharpe and Dohm Academic Scholarship at Birkbeck College, University of London, where he received his PhD in biophysics in 1991, under the supervision of Prof Tom L Blundell. He focused on development of methods for comparative modeling of protein three-dimensional structure and their implementation in the program MODELLER. He then went to the Department of Chemistry at Harvard University as a Jane Coffin Childs Memorial Fund postdoctoral fellow with Prof Martin Karplus, where he continued to develop comparative modeling methods and also studied simple lattice Monte Carlo models of protein folding. From 1995 to 2002, Dr Sali was first an assistant professor and then an associate professor at the Rockefeller University. In 2003, he moved to University of California at San Francisco as a Professor of Computational Biology in the Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research. He was a Sinsheimer Scholar, an Alfred P Sloan Research Fellow, and an Irma T Hirschl Trust Career Scientist. Dr Sali is an Editor of Structure and a Founder of Prospect Genomix, now Structural Genomix. He is interested in using computation grounded in the laws of physics and the theory of evolution to study the structure and function of proteins. He is aiming to improve and apply methods for (i) predicting the structures of proteins; (ii) determining the structures of macromolecular assemblies; and (iii) annotating the functions of proteins using their structures.

Comprehensive Medicinal Chemistry II
ISBN (set): 0-08-044513-6
ISBN (Volume 4) 0-08-044517-9; pp. 215–236