Chapter 1

Protein Structure Modeling

T. Schwede*,†, A. Sali‡, N. Eswar‡
and M. C. Peitsch§

1.1 Introduction

Knowledge of the three-dimensional (3D) structures of proteins provides invaluable insights into the molecular basis of their functions. Furthermore, the design of experiments aimed at understanding molecular mechanisms — such as site-directed mutagenesis, mapping of disease-related mutations, and the structure-based design of specific inhibitors — are greatly facilitated by the detailed knowledge of the spatial arrangement of key amino acid residues within the overall 3D structure. While great progress has been made in structure determination using experimental methods, such as X-ray crystallography (Chapter 22), high-resolution electron microscopy (Chapter 23) and

*Corresponding author.
†Swiss Institute of Bioinformatics, Biozentrum, University of Basel, Klingelbergstrasse 50/70, 4056 Basel, Switzerland. E-mail: torsten.schwede@unibas.ch.
‡Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, California Institute for Quantitative Biosciences (QB3), University of California at San Francisco, Byers Hall at Mission Bay, Suite 503B, 1700 4th Street, San Francisco, CA 94158-2330, USA.
§Novartis Institutes of BioMedical Research, Basel Klybeckstrasse 141, 4057 Basel, Switzerland.

nuclear magnetic resonance (NMR) spectroscopy (Chapter 24), these approaches are generally still expensive, time consuming, and not always applicable. Currently, less than 50 000 experimental protein structures have been released by the Protein Data Bank PDB[1] (Table 1.1), while another 3500 have been deposited but are still awaiting release. These structures correspond to approximately 17 000 different proteins (sharing less than 90% sequence identity among one another). Nevertheless, the number of structurally characterized proteins is small compared to the 300 000 annotated and curated protein sequences in the Swiss-Prot section of the UniProtKB[2] (http://www.expasy.org/sprot/). This number is even smaller when compared to the 5.2 million known protein sequences in the complete UniProtKB (October 2007). Even after removal of the highly redundant sequences from this database (above), the remaining 3.3 million sequences exceed the number of known 3D structures by more than two orders of magnitude. Thus, no experimental structure is available for the vast majority of protein sequences. This gap has widened over the last decade, despite the high-throughput X-ray crystallography pipelines developed for structural genomics.[3–5] Therefore, the gap in structural knowledge must be bridged by computation.

Computational methods for predicting the 3D structures of proteins enjoy a high degree of interest and are the focus of many

Table 1.1. Current PDB Holdings (October 2007)[a]

		Molecule Type				
		Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
Experimental Method	X-ray	36847	991	1709	24	39571
	NMR	5929	788	134	7	6858
	EM	106	11	40	0	157
	Other	83	4	4	2	93
	Total	42965	1794	1887	33	46679

[a]The content of the table was obtained from http://www.pdb.org (1). EM: electron microscopy.

research and service development efforts. The prediction of the 3D structure of a protein from its amino acid sequence remains a fundamental scientific problem and it is often considered as one of the *grand challenges* in computational biology and chemistry. Broadly, four different types of approaches are commonly in use. The first and most accurate approach is "comparative" or "homology" modeling that uses experimentally elucidated structures of related protein family members as templates to model the structure of the protein of interest (the "target"). These methods can only be employed when a detectable template of known structure is available. Second, fold recognition and threading methods are used to model proteins that have low or statistically insignificant sequence similarity to proteins of known structure (Chapter 2). Third, *de novo* (or *ab initio*) methods aim to predict the structure of a protein purely from its primary sequence, using principles of physics that govern protein folding and/or using information derived from known structures but without relying on any evolutionary relationship to known folds. Finally, a fourth group of methods, recently receiving a lot of attention, is the "integrative" or "hybrid" methods that combine information from a varied set of computational and experimental sources, including all those listed above.

1.2 Modeling Methods

1.2.1 *Comparative Protein Structure Modeling Techniques*

Template-based protein modeling techniques (aka "homology modeling" or "comparative modeling") exploit the evolutionary relationship between a target protein and templates with known experimental structures, based on the observation that evolutionarily related sequences generally have similar 3D structures. Most comparative modeling procedures consist of several consecutive steps, which can be repeated iteratively until a satisfactory model is obtained: 1) identification of suitable template structures related to the target protein and the alignment of the target and template(s) sequences; 2) modeling of

the structurally conserved regions and the prediction of structurally variable regions; 3) refinement of the initial model; and 4) evaluation of the resulting model(s).

1.2.1.1 *Identification of modeling templates and sequence alignments*

Identifying suitable template structures and calculating an accurate alignment of their sequences with that of the target are the key first steps of the comparative modeling process. The sequence identity of the target-template alignment is the most commonly used metric to quantify the similarity between the target and template(s) and is also a good predictor of the quality of the resulting model. It is thus crucial to consider the target-template sequence identity level when selecting template structures (Sections 1.2.2, 1.6 and Chapter 5), as this will have a critical impact on the quality of the resulting model and hence, its potential applications. The overall accuracy of models calculated from alignments with sequence identities of 40% or higher is almost always good (i.e. deviate by less than 2Å RMSD from the experimentally determined structure) (Section 1.2.2). As the target–template sequence identity falls below 30–40%, models that deviate significantly from the average accuracy are frequent (i.e. deviate by more than 2Å RMSD from an experimentally-determined structure). Alignment errors also tend to rapidly increase in this regime and become the most frequent cause of large errors in comparative models even when the correct template is chosen. Moreover, models based on alignments with such low sequence identities may have an entirely incorrect fold.[6]

While identifying and aligning sequences with similarities above 40% is relatively straightforward, more sensitive methods are needed for the lower levels of evolutionary relatedness between sequences. In recent years, significant progress has been made in the development of sensitive methods for sequence homology detection and alignment based on iterative profile searches, e.g. PSI-Blast,[7] Hidden Markov Models, e.g. SAM,[8] HMMER,[9] or profile-profile alignment such as FFAS03,[10] profile.scan,[11] and HHsearch.[12] Furthermore, in

the absence of a detectable sequence similarity, fold recognition and threading methods can be used to identify proteins with known structures, that share a common fold with the target sequence (Chapter 2).

1.2.1.2 *Generating all-atom models*

Comparative protein structure modeling yields an all-atom model of a protein, based on its alignment to one or more related template structures. Over the years, two commonly used approaches for model building have emerged and can be described as follows: the first is a rigid fragment assembly approach, in which an initial model is constructed from structurally conserved core regions of the template and from structural fragments obtained from either aligned or unrelated structures.[13,14] The initial model is then subjected to an optimization procedure to refine its geometry and stereochemistry (Section 1.2.1.3). The second approach relies on a single optimization strategy that attempts to maximize the satisfaction of spatial restraints obtained from the target-template alignment, known protein structures, and molecular mechanics force-fields.[15] Such an approach may not require a separate refinement step. However, most model building procedures are usually followed by the application of specialized protocols to enhance the accuracy of the non-conserved regions of the alignment such as loops[16,17] and/or side chains.[18,19]

1.2.1.3 *Model refinement*

Once an atomic model has been obtained, it can potentially be refined to idealize bond geometry and to remove unfavorable contacts that may have been introduced by the initial modeling process. The refinement will generally begin with an energy minimization step using one of the molecular mechanics force fields.[20,21] For further refinement, techniques such as molecular dynamics as well as Monte Carlo and genetic algorithm-based sampling methods[22–24] can be applied. For instance, in certain cases molecular dynamics has been reported to yield some improvement of side chain contacts and rotamer states.[25]

Monte Carlo sampling with focus on regions most likely to contain errors, while allowing the whole structure to relax in a physically realistic all-atom force field, can significantly improve the accuracy of models in terms of both the backbone conformations and the placement of core side chains.[26] Nevertheless, limitations still exist in sampling as well as force field accuracy.

1.2.1.4 *Model evaluation*

Model evaluation aims to recognize the various problems that might have occurred during the modeling process. Furthermore, estimating the overall geometrical accuracy of the individual regions of the model is an essential task of model evaluation. There are two kinds of evaluation schemes that are commonly employed. The first is "fold-assessment" that seeks to ensure the calculated models possess the correct fold and helps in detecting errors in template selection, fold recognition, and target-template alignment.[6,27-29] The second class of methods seeks to identify the model that is closest to the native structure out of a number of alternative models.[30-37] A combination of such assessments is usually employed to select the most accurate model from amongst a set of alternative models, generated based on different templates and/or alignments. In general, addressing these different types of assessment requires specialized scoring systems and classifiers (Chapters 3 and 4).

1.2.2 *Accuracy and Limitations of Comparative Protein Structure Modeling*

Comparative protein structure modeling relies on the evolutionary relationship between the target and template proteins. Consequently, the application of this approach is limited by 1) the availability of suitable template structures; 2) the ability of alignment methods to calculate an accurate alignment between the target and template sequences, even when the relationship between them is remote; and 3) the structural and functional divergence between the target and the template.[38]

The percentage of sequence identity between target and template correlates with model accuracy and often allows for a good first estimate of the model quality. As a rule of thumb, comparative models based on more than 50% sequence identity to their templates can be considered as "high accuracy models" and tend to have about 1 Å root mean square deviation (RMSD)[38] for the main-chain atoms, which is comparable to the accuracy of a medium-resolution NMR-derived structure or a low-resolution X-ray structure.[5,39] Inaccuracies are mainly found in the packing of side chains and loop regions. Comparative models based on 30 to 50% sequence identity can be considered "medium accuracy models", where the most frequent errors include side-chain packing errors, slight distortions of the protein core, inaccurate loop modeling, and sporadic alignment mistakes. Since alignment errors increase rapidly below 30% sequence identity and become the most substantial origin of errors in comparative models, comparative models based on less than 30% sequence identity are considered "low accuracy models".

1.2.2.1 *Template availability and structural diversity*

It has been observed that a very small number of different folds account for the majority of known structures,[40] and a recent study has argued that most sequences could already be modeled using known folds (or fragments of known folds) as templates.[41] Thus, for the majority of target protein domains, a structure with a similar fold would be available within the Protein Data Bank (PDB). However, models based on alignments with low sequence identity often provide accurate information only about the fold of the protein. As stated above, the accuracy of homology models decreases rapidly when the sequence identity between the target and template drops below 30%, mainly due to alignment errors and our inability to model structural differences between the target and the template. While the overall fold of proteins is often well conserved even at undetectable levels of sequence similarity, protein function — such as enzyme function and specificity — shows much higher variability,[42,43] even at high levels of sequence identity (above 50%). New methods

beyond simple homology-based assignments are therefore required for functional annotation of new genomic sequences, taking into account specific local structural features.

1.2.2.2 *Natively unstructured proteins*

Intrinsic disorder in proteins, i.e. the presence of unstructured regions, has been a focus of much attention recently, as it has been shown to be implicated in important biological roles, such as translation and transcriptional regulation, cell signaling, and molecular recognition in general. Several studies report examples of disordered proteins implicated in important cellular processes, undergoing transitions to more structured states upon binding to their target ligand, DNA, or other proteins.[44–46] New biological functions linked to native disorder are emerging, such as self-assembly of multi-protein complexes or involvement in RNA and protein chaperones.[47,48] Natively unstructured proteins pose a challenge for experimental structural determination as they can hinder the crystallization of proteins or interfere with NMR spectroscopy. Consequently, such proteins are also not amenable to modeling techniques, as it is unclear to what extent the "correct" conformation can be inferred by comparative modeling, as these protein regions depend on the context of a folded scaffold to assume a defined structure. However, computational approaches for detecting regions in protein sequences with a high propensity for intrinsic disorder have been successfully developed, based on the observation that such protein segments possess characteristic sequence properties.[49–52]

1.2.2.3 *Membrane proteins*

Membrane proteins are involved in a broad range of central cellular processes, including signaling and intercellular communication, vesicle trafficking, ion transport, and protein translocation. It is not surprising that the targets for ~40% of all therapeutic drugs in use today are human membrane proteins. These include targets such as ion channels, reuptake pumps as targets for anti-depressants, and the important group of 7-transmembrane G-protein coupled receptors

(GPCRs). However, membrane proteins pose formidable challenges to experimental structure determination by X-ray crystallography and NMR spectroscopy. Furthermore, human proteins often have no closely related homologs in prokaryotes or *archaea*, which would facilitate expression and crystallization. As a consequence, structures of membrane proteins are significantly underrepresented in the PDB. The 3D structures of only ~135 different membrane proteins are currently publicly available (1 January 2008). Consequently, prediction of membrane protein structures based on physical models that describe intra-protein and protein–solvent interactions in the membrane environment without relying on homologous template structures has been attempted by several groups.[53,54] An important challenge in the modeling of membrane protein structures is the presumed difference relative to the globular proteins. For example, it is believed that membrane proteins are "inside-out" globular proteins, with hydrophobic residues on the outside in contact with the lipid bilayer and polar residues on the inside in the protein core. This design may render the standard scoring functions used for the modeling of globular proteins less suitable for use with membrane proteins. Most recently, a new scoring function was developed in Rosetta to account for such differences.[55]

1.2.3 *De novo Modeling Techniques*

Comparative protein structure modeling methods are only able to produce highly accurate models for protein sequences for which sufficient template information is available on the structures of homologous proteins. However, these methods are not suited to predict parts of sequences that are not aligned with the template sequences, e.g. long variable loop regions, or completely novel folds that have not been observed before. In contrast, *de novo* modeling methods do not explicitly rely on whole known structures as templates. Thus, the structure of any protein can be predicted by these *de novo* methods.

The term *ab initio* prediction often refers to the subset of *de novo* methods that rely on energy functions based solely on physicochemical interactions, not on the PDB. Such approaches, using full-atom simulations with empirical force fields as well as explicit and implicit

solvent models, have been successful in predicting the folding of short peptides[56,57] and in discriminating between the native state and a static set of decoys.[58] However, from a practical protein structure prediction perspective, there are still limitations with regards to protein size and accuracy of the predictions.

Most of the successful *de novo* prediction methods that are applicable to larger protein segments (up to ~150 residues) use information from known protein structures.[59] *De novo* methods assume that the native state of a protein is at the global free energy minimum and carry out a large-scale search of conformational space for protein tertiary structures that are particularly low in free energy for the given amino acid sequence. The working hypothesis of this approach is that local amino acid sequence propensities bias each local segment of a polypeptide chain towards a small number of alternative local structures and that non-local interactions preferentially stabilize native-like arrangements of these otherwise transitory local structures. For example, the Rosetta method developed by Baker and coworkers uses an ensemble of short structural fragments extracted from the PDB.[60] These fragments are then assembled in a Monte Carlo search strategy using a scoring function that favors non-local properties of native protein structures such as hydrophobic burial, compactness, and pairing of β-strands.[22,60,61] Using fragments of known structures ensures that the local interactions are close to optimal, thereby reducing the demand on the free energy function. The Rosetta fragment assembly strategy has been successfully applied to *de novo* structure prediction, as well as to modeling of structurally variable regions (loops, insertions) in comparative protein structure models.

The TASSER (Threading/ASSEmbly/Refinement) method developed by Skolnick, Zhang and coworkers uses tertiary restraints derived from threading results to restrict the conformational search space. The query sequence is first threaded through the structures representative of the PDB to identify appropriate local fragments for further structural reassembly. For a given alignment, an initial full-length model is built by connecting the continuous secondary structure fragments through a random walk, followed by parallel-exchange Monte Carlo sampling for refinement.[62,63]

De novo modeling techniques have made tremendous progress over the last decade, and several individual examples of highly accurate predictions have been reported. However, there are still significant limitations that restrict their application for routine use: the computational demand is immense and therefore limits these methods to relatively small systems. In parallel, the overall quality of the resulting models decreases with the increasing size of the protein. As a result, the accuracy of *de novo* predictions is in general still poor, despite a number of positive examples. In CASP7 (Section 1.5), it was generally not possible to correctly predict the overall fold for a majority of the *de novo* modeling targets.[64]

1.3 Protein Modeling and Structural Genomics

Comparative protein structure modeling and experimental protein structure determination complement each other, with the long-term goal of making three-dimensional atomic-level information of most proteins obtainable from their corresponding amino acid sequences. To achieve structural coverage of a majority of sequenced genes, systematic sampling of major protein families with experimental protein structures is essential (unless the *de novo* methods become perfect). Structural genomics is a worldwide initiative aimed at rapidly determining a large number of protein structures using X-ray crystallography and NMR spectroscopy in a high-throughput mode.[65,66] As a result of concerted efforts in technology and methodology development in recent years, each step of experimental structure determination has become more efficient, less expensive, and more likely to succeed.[67] Structural genomics initiatives are making significant contribution to both the scope and depth of our structural knowledge about protein families. Although worldwide structural genomics initiatives only account for ~20% of the new structures, these contribute approximately to three quarters of the new structurally characterized families and over five times as many novel folds as classical structural biology.[68–73]

Most structural genomics consortia follow specific objectives that include focusing on certain protein classes, such as membrane

proteins, protein families with special biomedical relevance, enlarging the coverage of sequence space on the domain level, and determining all the proteins in a model genome. They are applying sophisticated bioinformatics strategies for target selection to maximize the gain in novel insights into protein function from a structural perspective.[68,70,71,74–76]

In the light of the ever-growing amount of genome sequencing data, the structure of most of the proteins, even with structural genomics, will be modeled and not elucidated experimentally. From a modeling-centric perspective, the selection of structural genomics targets should thus be such that most of the remaining sequences can be modeled with useful accuracy by comparative modeling. As discussed before, the accuracy of the comparative models currently declines sharply below the 30% sequence identity. Thus, target selection strategies should aim at systematic sampling of protein structures to ensure that most of the remaining sequences are related to at least one experimentally elucidated structure at more than the 30% sequence identity.[5] Using this cutoff, it has been estimated that a minimum of 16 000 targets must be determined to cover 90% of all the protein domain families, including those of membrane proteins.[77] Such estimates show large variations, depending on the level of sequence identity that is assumed to ensure sufficiently accurate model building, and how this coverage is calculated. Recently, it has been proposed to reduce this number to a manageable size by prioritizing structurally uncharacterized protein families from PFAM according to the number of family-members.[78] However, it has been argued that such coarse-grained target selection is suboptimal in terms of reliable structural and functional annotation, and a selection of "fine-grain" targets from within larger coarse-grained families of distantly related proteins would be required to provide a more thorough coverage of functional space as it relates to protein structure.[68]

Until recently, sequence databases were highly biased towards proteins of known function from a relatively small set of model organisms, a result of targeted protein sequencing. However, in the last decade, whole-genome sequencing efforts have presumably reduced or eliminated this bias. We are, however, on the threshold of a new dimension in sequence diversity. The recent meta-genomics projects

(which are based on shotgun sequencing of populations of microorganisms) have yielded new insights into the distribution of (mainly microbial) protein families. As there is an approximately linear relationship between the number of sequence clusters and the number of protein sequences, this indicates that there remain many more protein families to be discovered. This, in turn, has direct implications on the selection of targets for structural genomics.[79]

1.4 Integrative (Hybrid) Modeling Techniques

Biological function is seldom effected by a single protein molecule in isolation. It is most often the result of transient or stable interactions among individual proteins in the cell. Most of these interactions remain uncharacterized by traditional structural biology techniques such as X-ray crystallography (Chapter 22) and NMR spectroscopy (Chapter 24). This gap is being bridged by several emerging experimental approaches that vary in terms of the information they provide.[80] For example, the stoichiometry and composition of protein components in an assembly can be determined by methods such as quantitative immunoblotting and mass spectrometry. The shape of the assembly can be revealed by electron microscopy and small angle X-ray scattering. The positions of the components can be elucidated by cryoelectron microscopy and labeling techniques. Whether or not components interact with each other can be measured by mass spectrometry, yeast two-hybrid and affinity purification. The relative orientations of the components and information about interacting residues can be inferred from cryoelectron microscopy, hydrogen/deuterium exchange, hydroxyl radical footprinting, and chemical-crosslinking.

When the approaches dominated by a single source of information fail, simultaneous consideration of all the available information about the composition and structure of a given assembly, irrespective of its source, can sometimes be sufficient to calculate a useful structural model. Thus, integrative modeling methods convert the experimental data derived from the methods listed above into a structural model of a macromolecular assembly through computation[80] (Fig. 1.1). Such an approach can be used to uncover

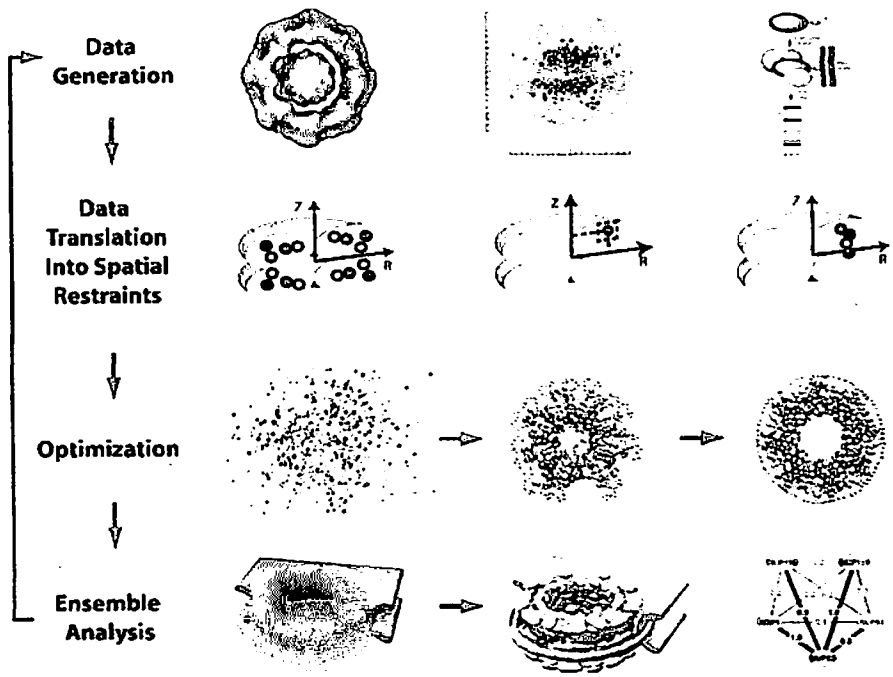

Fig. 1.1 Integrative structure determination. The four steps of determining a structure by integration of varied data are illustrated with the example of the nuclear pore complex.[80,84,132] First, structural data are generated by experiments, such as electron microscopy (*left panel*), immunoelectron microscopy (*middle panel*), and affinity purification of subcomplexes (*right panel*); many other types of information can also be added. Second, the data and theoretical considerations are expressed as spatial restraints ensuring the observed symmetry and shape of the assembly (electron microscopy, *left panel*), positions of constituent gold-labeled proteins (immunoelectron microscopy, *middle panel*), and proximity among the constituent proteins (affinity co-purification, *right panel*). Third, an ensemble of structural solutions that satisfy the data is obtained by minimizing the violations of the spatial restraints (from *left* to *right*). Fourth, the ensemble is clustered into sets of distinct solutions (*left panel*) as well as analyzed in different representations, such as protein positions (*middle panel*) and protein-protein contacts (*right panel*). The integrative approach to structure determination has several advantages: (i) it benefits from the synergy among the input data, minimizing the drawback of incomplete, inaccurate, and/or imprecise data sets (although each individual restraint may contain little structural information, the concurrent satisfaction of all restraints derived from independent experiments may drastically reduce the degeneracy of structural solutions); (ii) it can potentially

the molecular architecture of macromolecular assemblies and even atomic models of protein complexes. Even when this model is of relatively low resolution and accuracy, it can still be helpful for studying the function and evolution of the corresponding assembly; it also provides the necessary starting point for a higher resolution study.

An example of a simple hybrid approach is building a pseudo-atomic model of a large assembly by fitting atomic structures of subunits into its cryoelectron microscopy map.[81] Unassigned or partially assigned NMR spectroscopy data and fragment-based modeling approaches have been combined to improve structure refinement in terms of its accuracy, efficiency, and success rate.[82,83] A variety of different types of information, such as symmetry and protein proximity, have been used to characterize large symmetrical assemblies, including the nuclear pore complex,[84,85] EscJ from the type III secretion system,[86] and the AAA+ ring complexes.[87]

1.5 Assessment and Evaluation of Prediction Accuracy

Protein structure modeling is maturing and therefore widely used as a scientific research tool today. Consequently, it is increasingly important to evaluate to what extent the current prediction methods meet the accuracy and requirements of different scientific applications (Chapter 5). A good way to assess the reliability of different protein structure modeling methods *a posteriori* is by evaluating the results of blind predictions after the corresponding protein structures have been determined experimentally. One such effort is the biannual "Community Wide Experiment on the Critical Assessment of

produce all structures that are consistent with the data, not just one; (iii) the variation among the structures consistent with the data allows us to assess the sufficiency of the data and the precision of the representative structure; (iv) it can make the process of structure determination more efficient by indicating what measurements would be the most informative. (This figure was reproduced from Fig. 5 in Ref. 80).

Techniques for Protein Structure Prediction" (CASP).[88,89] During a CASP trial, research groups apply their prediction methods to sequences for which the experimental structure is about to be determined. The accuracy of these blind predictions is then assessed independently once the structures are made available. There are also web servers, LIVEBENCH[90] and EVA,[91] that assess protein structure prediction servers on an automated and continuous basis using sequences from the PDB, before their structures are released, as modeling targets.

1.5.1 *Critical Assessment of Techniques for Protein Structure Prediction (CASP)*

The biannual CASP experiments aim to assess the progress of protein structure prediction methods.[88,92] Besides using classical measures for assessing the accuracy of the Cα positions of the models, several additional criteria were introduced in CASP7 to ensure that the assessment appraises the overall quality of the models, as well as those features of the predictions that are relevant to their usefulness in specific scientific applications, such as the fraction of correctly modeled hydrogen bond interactions (HBscore), the suitability of models for phasing X-ray diffraction data, assessment of the accuracy of predicted cofactor binding sites, and accuracy of the model error estimates provided by the predictors.

In the latest edition of CASP (round 7 in 2006),[39,64,89,93] the general trends observed in the previous years continued: comparative modeling remained by far the most accurate technique for protein structure modeling. However, the majority of predictions submitted in the category of template-based modeling (TBM) were again closer to the template than to the real structure, and only in a few cases, some improvement over a model based on a single best template structure was observed. The fact that no group would outperform a virtual predictor submitting models based on the single best template for each target indicates that template identification and alignment are by no means solved problems and constitute a major bottleneck, besides the challenging question of model refinement. Impressively,

successful refinement of model coordinates to a value closer to the experimental structure has been observed, at least in a small number of cases.[22,94]

One of the most remarkable results of CASP7 was that automated prediction servers have matured significantly in the recent years: six of the top 25 groups in the assessment of template-based models were predictors using automated prediction servers, which produce their models without manual intervention. In 29% of a total of 108 cases, the best model for an individual prediction target was submitted by a server. The best prediction server[63] was ranked third over all, i.e. it outperformed all but two of the participating groups.[93,94]

1.5.2 *EVA-CM — Continuous Automated Assessment of Prediction Servers*

The goal of EVA[91] is to evaluate the sustained performance of protein structure prediction servers through objective measures for prediction accuracy in a fully automated manner. Every week, test sequences are automatically submitted to prediction servers and the results are evaluated and posted on the EVA web sites, thereby providing a continuous, fully automatic and statistically significant analysis of structure prediction servers. Besides comparative modeling, EVA assesses the prediction of secondary structure, inter-residue distances and contacts, and threading.

1.5.3 *Model Quality Evaluation*

Retrospective assessment of the average accuracy of individual modeling methods via projects such as CASP or EVA is invaluable for the development of modeling techniques, but unfortunately does not allow drawing of any conclusions about the accuracy of a specific model, as the correct answer is unknown in a real-life situation. Since the usefulness of predictions crucially depends on their accuracy, a means of reliably predicting the likely accuracy of a protein structure model in the absence of its known 3D structure is an important problem in protein structure prediction (Section 1.2.1.4). Accurate

estimates of the errors in a model are an essential component of any predictive method — protein structure prediction not being an exception.

Different scoring schemes have been developed to determine whether or not a model has the correct fold, to differentiate between the native and near-native states, to select the most near-native model in a set of decoys, and to provide quantitative estimates for the coordinate error of the predicted amino acids (Section 1.2.1.4). A variety of methods have been applied to address these tasks, such as physics-based energies, knowledge-based potentials (Chapter 3), combined scoring functions, and clustering approaches. Combined scoring functions integrate several different scores, aiming to extract the most informative features from each of the individual input scores (Chapter 4). Clustering approaches use consensus information from an ensemble of protein structure models provided by different methods.

1.6 Application of Protein Models

1.6.1 *Typical Applications of Protein Models*

The suitability of protein models for specific applications crucially depends on their accuracy. There is a wide range of applications for comparative models, such as designing experiments for site-directed mutagenesis or protein engineering, predicting ligand binding sites and docking small molecules in structure-based drug discovery,[95,96] studying the effect of mutations and SNPs,[97,98] phasing X-ray diffraction data in molecular replacement,[26,99] as well as protein engineering and design.[100] See Chapter 5 for a more detailed discussion about applications of models.

Although the target-template sequence identity generally correlates well with the overall model accuracy, it is often not suitable for making decisions about the usability of models for specific applications. There is a need for new measures to come up with more reliable estimates of model quality. For instance, applications in drug design require a very high accuracy of the local sidechain positions in the binding site, much more so that the overall global accuracy of the backbone.[94,101] Local estimates of the expected model accuracy on a

per residue or per atom level would be crucial for many applications, e.g. phasing of crystallographic diffraction data.[39]

1.6.2 *Modeling GPCRs*

Modeling G-protein-coupled receptors has drawn much attention due to their relevance as drug targets. Constraints-based and homology modeling[102-104] has been used as a tool to obtain structural models for GPCRs, at first based on the structures of bacteriorhodopsin,[105,106] and since 2000 using the high resolution X-ray structure of bovine rhodopsin[107] as a template for modeling.[108,109] Only recently the first structure of a GPCR bound to a diffusible ligand, the human β_2-adrenergic G-protein coupled receptor,[110,111] has become available and may now serve as a more suitable template for modeling other members of the class A GPCRs. However, the level of sequence identity within the members of the class A GPCRs is often very low, seriously limiting the accuracy of the local alignment. Especially the conformations of non-conserved inter-helical loops are difficult to model using comparative techniques. Retrospectively, we can analyze the accuracy of the "historic" comparative models built for the human β_2-adrenergic receptor based on the rhodopsin structure as templates. While the overall arrangement of the 7 trans-membrane helix segments is generally correctly represented, significant differences are observed in the relative orientation and shifts of the helices with regard to the center of the receptor (Fig. 1.2).

The ligand-binding pocket is, with regard to rhodopsin, formed by both the structurally conserved and divergent segments. Most deviations are observed for helices III, V, and the extracellular loop ECL2, which connects helices IV and V (Fig. 1.2). While ECL2 is forming a β-sheet structure in rhodopsin, in β_2-adrenergic receptor it contains an unexpected additional α-helical segment and a second disulfide bridge that might stabilize the more solvent exposed conformation. Consequently, specific interactions between the ligand molecule and side chains forming the binding pocket are only partially reproduced by a comparative model based on rhodopsin (Fig. 1.3). See Refs. 110, 111 for a detailed discussion of the individual structural differences, as well as discussion of the activation mechanism.

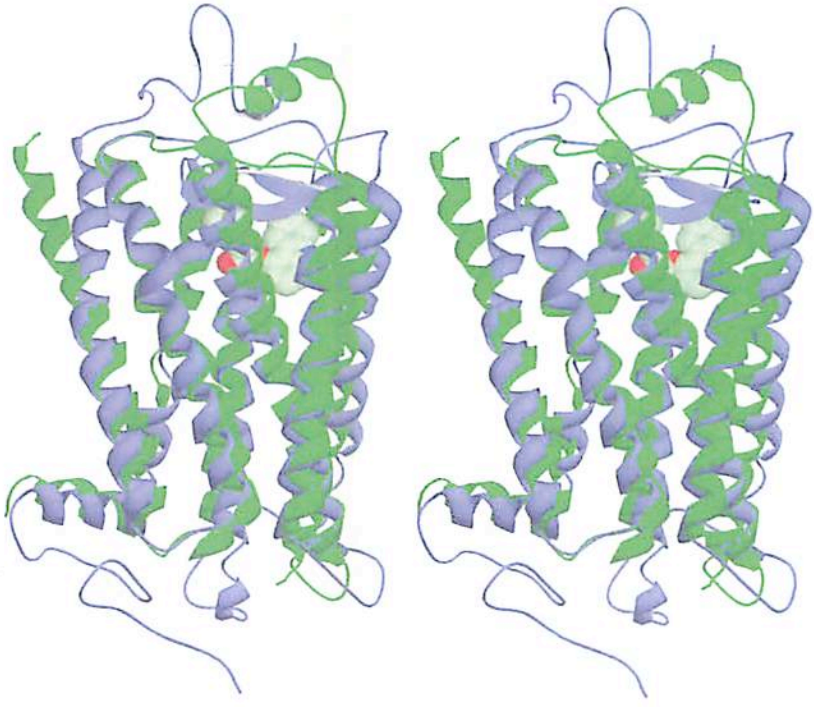

Fig. 1.2 Ribbon representation of the human β_2-adrenergic G-protein coupled receptor with bound ligand carazolol (green, PDB: 2rh1[110]) and the bovine rhodopsin (blue, PDB: 1u19[107]). Bovine rhodopsin has been the only available high resolution template for modeling class A GPCRs until the structure of β_2-adrenergic receptor has been solved in 2007. (Superposition, stereo view).

1.7 Major Protein Modeling Resources

1.7.1 *Protein Modeling Servers and Software Tools*

The huge and constantly growing number of structurally uncharacterized protein sequences, together with the increasing number of available template structures requires the development of automated, stable and reliable modeling methods. Modeling of protein structures usually requires expertise in structural biology and the use of highly specialized computer programs for each of the individual steps of the modeling process. Therefore, automated modeling pipelines

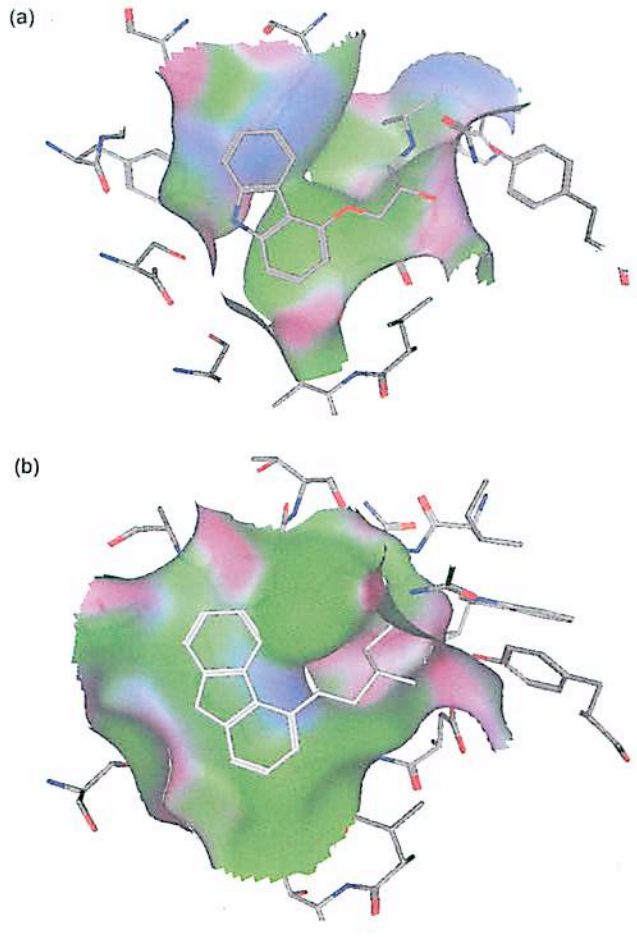

Fig. 1.3 The ligand binding site of the β_2-adrenergic G-protein coupled receptor. The experimentally elucidated structure in panel (a) (PDB: 2rh1[110]) as compared to the comparative model based on bovine rhodopsin as template in planel (b) (PDB: 1u19[107]).

with integrated expert knowledge such as SWISS-MODEL[14,112–114] and MODPIPE[15,115] were established 15 years ago and have been successfully applied to large data sets.[3,116–120]

Today, there is a plethora of modeling services available on the Internet. Therefore, the question is what is the most appropriate method for a specific target? Meta-servers — methods that use the

Table 1.2. List of Protein Modeling Servers and Software. For a more exhaustive list, see Refs. 93 and 122

	Modeling Server
SwissModel[112–114,124]	http://swissmodel.expasy.org
ModWeb[115]	http://salilab.org/modweb/
I-Tasser[63]	http://zhang.bioinformatics.ku.edu/I-TASSER/
Robetta[133]	http://robetta.bakerlab.org
	Software Tools
HHPred[134]	http://toolkit.tuebingen.mpg.de/hhpred
Modeller[15,115]	http://salilab.org/modeller/
SCWRL3[19]	http://dunbrack.fccc.edu/SCWRL3.php
WhatIf[135]	http://swift.cmbi.ru.nl/whatif/
Rosetta[60]	http://www.rosettacommons.org

results of other servers as input to generate their predictions — are aiming to address this question.[90,121] The general opinion in the community has been that the models generated using a combination of automated predictions and human expertise are superior to those generated using purely automated servers.[90] However, it appears that this view might have to be revised in the near future as the gap between human predictors and servers is closing. Table 1.2 provides examples of the major available resources; see Refs. 93, 122 for a more comprehensive list.

1.7.2 *Protein Model Databases*

Depositions to the PDB are restricted to atomic coordinates that are substantially determined by experimental measurements on specimens containing biological macromolecules.[123]

Currently, the PDB holds approximately 50 000 entries representing 17 000 different proteins. Using these experimentally elucidated structures as templates, several millions of comparative protein models have been generated for the protein sequences contained in the UniProtKB database.[3,4,124,125] Databases of annotated comparative models increase the efficiency for expert users, allow cross-referencing with other (non-structure-centric) resources, and make

Table 1.3. Databases of Automated Comparative Protein Models

Model Database Resources		Refs.
MODBASE	http://www.salilab.org/modbase/	125, 126
SWISS-MODEL Repository	http://swissmodel.expasy.org/repository/	117, 120, 124
Protein Model Portal	http://www.proteinmodelportal.org	

Table 1.4 Protease Models for Entries referenced in the MEROPS Database available in the Protein Model Portal

Group	Number of UniProtKB Entries	Number of Models	Average Sequence Identity with Best Template
Grand Total	6869	28701	39.0%
SWISS-MODEL Repository	3362	5440	69.9%
MODBASE	5001	21471	33.2%
CSMP (Center for Structures of Membrane Proteins)	7	17	19.9%
MCSG (Midwest Center for Structural Genomics)	48	48	28.2%
NESG (Northeast Center for Structural Genomics)	199	244	17.7%
NYSGXRC (New York SGX Center for Structural Genomics)	748	1481	16.9%
PDB[a]	400	2338	N.A.
Protease sequences without structure or model	1342	0	N.A.

[a]Experimentally elucidated protease structures. N.A., not applicable.

comparative models accessible to non-experts. Many specialized efforts exist for specific protein families, or specific organisms. These resources are often manually curated, which poses challenges in terms of maintaining a reasonable update frequency when new template structures and new or updated sequence information become available. Generic model databases such as MODBASE[125,126] and the

SWISS-MODEL Repository[120,124] apply entirely automated techniques for large-scale comparative protein structure modeling.

The Protein Model Portal (http://www.proteinmodelportal.org) has recently been developed as part of the PSI Structural Genomics Knowledge Base to provide an integrated access to the various databases containing structural information and thereby implementing the first step of the community workshop recommendation[123] on archiving structural models of biological macromolecules. Currently, automatically-derived models from six structural genomics centers, MODBASE and SWISS-MODEL Repository are accessible through a single search interface. As an example, we have analyzed all the protease families referenced in the MEROPS database[127] for the number of protein models — and their average sequence identity to the best modeling template — currently available from the Protein Model Portal. It is interesting to note that even in this highly studied class of proteins, there is no structural information available, experimental or modeled, for approximately 20% of the sequences in UniProtKB.

1.8 Future Outlook

1.8.1 *Model Refinement*

Comparative protein structure modeling has matured over the last decade and is now routinely used in many practical applications. There has been a continuous increase in the overall accuracy of protein structure models due to progress in the quality of the sequence-structure alignments as well as the increased availability of high quality template structures. However, comparatively little progress has been made in refining the initial models away from the template closer to the target structure. Model refinement is particularly relevant for models based on alignments with a sequence identity below 30%, which is the typical situation in comparative modeling. Many biomedical applications (Section 1.6) are critically dependent on model accuracy, and the accuracy achieved by comparative modeling based on low sequence identity templates is often insufficient. Improving

the accuracy of comparative models beyond the information derived from the template therefore continues to be one of the key questions in the future. Although examples of successful model refinement using molecular dynamics methods have been described occasionally, these methods do not seem to be generally successful.[25,128] The challenges with refinement seem to reside in the limitations of the currently available force fields (which do not accurately represent the energetic interactions of the native state of the protein structure), as well as in the computational effort required for sampling a highly dimensional and rugged energy landscape, which is necessary to identify the global minimum.[22,23,26,129]

1.8.2 *Integrative (Hybrid) Modeling*

Cryoelectron microscopy is emerging as a key technique for studying 3D structures of multi-component macromolecular complexes with masses >250 kDa, such as membrane proteins, cytoskeletal complexes, ribosomes, quasi spherical viruses, molecular chaperones, flagella, ion channels, and oligomeric enzymes. Electron cryotomography even enables the observation of macromolecules inside a living cell in its native state.[130] Various modeling approaches are being developed that utilize cryoelectron microscopy density maps as a constraint in deriving a pseudo-atomic model of the molecular components within a larger complex. Because of the significant likelihood of conformational differences between isolated domains and biological assemblies, additional research leading to the development of reliable hybrid modeling methods, which are able to correctly include structural information from various experimental sources of different resolution and reliability, is essential. The important structural information from hybrid models, generating a synoptic image of the heterogeneous information available for a given macromolecular system, is expected to increase sharply in the coming years. Naturally, this raises the question of whether it will be feasible at one point to combine all these data, together with other data related to the overall cellular structure, to construct a quantitative spatial and temporal model of the cell.[131]

References

1. Berman HM, Westbrook J, Feng Z, *et al.* (2000) The Protein Data Bank. *Nucl Acids Res* **28**(1): 235–242.
2. Bairoch A, Apweiler R, Wu CH, *et al.* (2005) The Universal Protein Resource (UniProt). *Nucl Acids Res* **33**(Database Issue): D154–159.
3. Peitsch MC. (1997) Large scale protein modelling and model repository. *Proc Int Conf Intell Syst Mol Biol* **5**: 234–236.
4. Peitsch MC, Schwede T, Guex N. (2000) Automated protein modelling — the proteome in 3D. *Pharmacogenomics* **1**(3): 257–266.
5. Baker D, Sali A. (2001) Protein structure prediction and structural genomics. *Science* **294**(5540): 93–96.
6. Melo F, Sali A. (2007) Fold assessment for comparative protein structure modeling. *Protein Sci* **16**(11): 2412–2426.
7. Altschul SF, Madden TL, Schaffer AA, *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl Acids Res* **25**(17): 3389–3402.
8. Karplus K, Barrett C, Hughey R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**(10): 846–856.
9. Eddy SR. (1998) Profile hidden Markov models. *Bioinformatics* **14**(9): 755–763.
10. Jaroszewski L, Rychlewski L, Li Z, *et al.* (2005) FFAS03: a server for profile – profile sequence alignments. *Nucl Acids Res* **33**(Web Server Issue): W284–288.
11. Marti-Renom MA, Madhusudhan MS, Sali A. (2004) Alignment of protein sequences by their profiles. *Protein Sci* **13**(4): 1071–1087.
12. Soding J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**(7): 951–960.
13. Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM. (1987) Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326**(6111): 347–352.
14. Peitsch MC, Jongeneel CV. (1993) A 3-D model for the CD40 ligand predicts that it is a compact trimer similar to the tumor necrosis factors. *Int Immunol* **5**(2): 233–238.
15. Sali A, Blundell TL. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**(3): 779–815.
16. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B. (2007) Loop modeling: sampling, filtering, and scoring. *Proteins.*
17. Jacobson MP, Pincus DL, Rapp CS, *et al.* (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins* **55**(2): 351–367.
18. Lovell SC, Word JM, Richardson JS, Richardson DC. (2000) The penultimate rotamer library. *Proteins* **40**(3): 389–408.
19. Canutescu AA, Shelenkov AA, Dunbrack RL, Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* **12**(9): 2001–2014.

20. Brooks BR, Bruccoleri RE, Olafson BD, *et al.* (1983) CHARMM: a program for macromolecular energy, minmimization, and dynamics calculations. *J Comput Chem* 4: 187–217.
21. Cornell WD, Cieplak P, Bayly CI, *et al.* (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* 117: 5179–5197.
22. Das R, Qian B, Raman S, *et al.* (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* 69(S8): 118–128.
23. Han R, Leo-Macias A, Zerbino D, *et al.* (2007) An efficient conformational sampling method for homology modeling. *Proteins* 10.1002/prot.21672.
24. Qian B, Ortiz AR, Baker D. (2004) Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. *Proc Natl Acad Sci USA* 101(43): 15346–15351.
25. Chen J, Brooks CL, 3rd. (2007) Can molecular dynamics simulations provide high-resolution refinement of protein structure? *Proteins* 67(4): 922–930.
26. Qian B, Raman S, Das R, *et al.* (2007) High-resolution structure prediction and the crystallographic phase problem. *Nature.*
27. Domingues FS, Koppensteiner WA, Jaritz M, *et al.* (1999) Sustained performance of knowledge-based potentials in fold recognition. *Proteins* (3): 112–120.
28. McGuffin LJ, Jones DT. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* 19(7): 874–881.
29. Miyazawa S, Jernigan RL. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 256(3): 623–644.
30. Lazaridis T, Karplus M. (1999) Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol* 288(3): 477–487.
31. Gatchell DW, Dennis S, Vajda S. (2000) Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins* 41(4): 518–534.
32. Vorobjev YN, Hermans J. (2001) Free energies of protein decoys provide insight into determinants of protein stability. *Protein Sci* 10(12): 2498–2506.
33. Seok C, Rosen JB, Chodera JD, Dill KA. (2003) MOPED: Method for optimizing physical energy parameters using decoys. *J Comput Chem* 24(1): 89–97.
34. Tsai J, Bonneau R, Morozov AV, *et al.* (2003) An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins* 53(1): 76–87.
35. Zhu J, Zhu Q, Shi Y, Liu H. (2003) How well can we predict native contacts in proteins based on decoy structures and their energies? *Proteins* 52(4): 598–608.
36. Eramian D, Shen MY, Devos D, *et al.* (2006) A composite score for predicting errors in protein structure models. *Protein Sci* 15(7): 1653–1666.

37. Shen MY, Sali A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15**(11): 2507–2524.
38. Chothia C, Lesk AM. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**(4): 823–826.
39. Read RJ, Chavali G. (2007) Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins* **69**(S8): 27–37.
40. Orengo CA, Thornton JM. (2005) Protein families and their evolution-a structural perspective. *Ann Rev Biochem* **74**: 867–900.
41. Zhang Y, Skolnick J. (2005) The protein structure prediction problem could be solved using the current PDB library. *Proc Natl Acad Sci USA* **102**(4): 1029–1034.
42. Rost B. (2002) Enzyme function less conserved than anticipated. *J Mol Biol* **318**(2): 595–608.
43. Tian W, Skolnick J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* **333**(4): 863–882.
44. Dyson HJ, Wright PE. (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**(3): 197–208.
45. Radivojac P, Iakoucheva LM, Oldfield CJ, *et al.* (2007) Intrinsic disorder and functional proteomics. *Biophys J* **92**(5): 1439–1456.
46. Fink AL. (2005) Natively unfolded proteins. *Curr Opin Struct Biol* **15**(1): 35–41.
47. Namba K. (2001) Roles of partly unfolded conformations in macromolecular self-assembly. *Genes Cells* **6**(1): 1–12.
48. Tompa P, Csermely P. (2004) The role of structural disorder in the function of RNA and protein chaperones. *FASEB J* **18**(11): 1169–1175.
49. Bordoli L, Kiefer F, Schwede T. (2007) Assessment of disorder predictions in CASP7. *Proteins* **69**(S8): 129–136.
50. Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **61**(7): 176–182.
51. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**(13): 2138–2139.
52. Schlessinger A, Liu J, Rost B. (2007) Natively unstructured loops differ from other loops. *PLoS Comput Biol* **3**(7): e140.
53. Barth P, Schonbrun J, Baker D. (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc Natl Acad Sci USA* **104**(40): 15682–15687.
54. Zhang Y, Devries ME, Skolnick J. (2006) Structure modeling of all identified G protein-coupled receptors in the human genome. *PLoS Comput Biol* **2**(2): e13.

55. Yarov-Yarovoy V, Schonbrun J, Baker D. (2006) Multipass membrane protein structure prediction using Rosetta. *Proteins* **62**(4): 1010–1025.
56. Jayachandran G, Vishal V, Garcia AE, Pande VS. (2007) Local structure formation in simulations of two small proteins. *J Struct Biol* **157**(3): 491–499.
57. Muff S, Caflisch A. (2007) Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a beta-sheet miniprotein. *Proteins.* 10.1002/prot.21565.
58. Verma A, Wenzel W. (2007) Protein structure prediction by all-atom free-energy refinement. *BMC Struct Biol* **7**: 12.
59. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. (2007) The protein folding problem: when will it be solved? *Curr Opin Struct Biol* **17**(3): 342–346.
60. Rohl CA, Strauss CE, Misura KM, Baker D. (2004) Protein structure prediction using Rosetta. *Meth Enzymol* **383**: 66–93.
61. Rohl CA, Strauss CE, Chivian D, Baker D. (2004) Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55**(3): 656–677.
62. Wu S, Skolnick J, Zhang Y. (2007) *Ab initio* modeling of small proteins by iterative TASSER simulations. *BMC Biol* **5**: 17.
63. Zhang Y. (2007) Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69**(S8): 108–117.
64. Jauch R, Yeo HC, Kolatkar PR, Clarke ND. (2007) Assessment of CASP7 structure predictions for template free targets. *Proteins* **69**(S8): 57–67.
65. Burley SK. (2000) An overview of structural genomics. *Nat Struct Biol* **7**(Suppl): 932–934.
66. Thornton J. (2001) Structural genomics takes off. *Trends Biochem Sci* **26**(2): 88–89.
67. Slabinski L, Jaroszewski L, Rodrigues AP, *et al.* (2007) The challenge of protein structure determination — lessons from structural genomics. *Protein Sci* **16**(11): 2472–2482.
68. Marsden RL, Lewis TA, Orengo CA. (2007) Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. *BMC Bioinform* **8**: 86.
69. Todd AE, Marsden RL, Thornton JM, Orengo CA. (2005) Progress of structural genomics initiatives: an analysis of solved target structures. *J Mol Biol* **348**(5): 1235–1260.
70. Chandonia JM, Brenner SE. (2006) The impact of structural genomics: expectations and outcomes. *Science* **311**(5759): 347–351.
71. Liu J, Montelione GT, Rost B. (2007) Novel leverage of structural genomics. *Nat Biotechnol* **25**(8): 849–851.

72. Gileadi O, Knapp S, Lee WH, *et al.* (2007) The scientific impact of the Structural Genomics Consortium: a protein family and ligand-centered approach to medically-relevant human proteins. *J Struct Funct Genomics* **8**: 107–119.
73. Levitt M. (2007) Growth of novel protein structural data. *Proc Natl Acad Sci USA* **104**(9): 3183–3188.
74. Chen L, Oughtred R, Berman HM, Westbrook J. (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics* **20**(16): 2860–2862.
75. Liu J, Hegyi H, Acton TB, *et al.* (2004) Automatic target selection for structural genomics on eukaryotes. *Proteins* **56**(2): 188–200.
76. Bussow K, Scheich C, Sievert V, *et al.* (2005) Structural genomics of human proteins — target selection and generation of a public catalogue of expression clones. *Microb Cell Fact* **4**: 21.
77. Vitkup D, Melamud E, Moult J, Sander C. (2001) Completeness in structural genomics. *Nat Struct Biol* **8**(6): 559–566.
78. Chandonia JM, Brenner SE. (2005) Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. *Proteins* **58**(1): 166–179.
79. Yooseph S, Sutton G, Rusch DB, *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**(3): e16.
80. Robinson CV, Sali A, Baumeister W. (2007) Molecular sociology of the cell. *Nature* **450**: 973–982.
81. Topf M, Baker ML, Marti-Renom MA, *et al.* (2006) Refinement of protein structures by iterative comparative modeling and CryoEM density fitting. *J Mol Biol* **357**(5): 1655–1668.
82. Lee SY, Zhang Y, Skolnick J. (2006) TASSER-based refinement of NMR structures. *Proteins* **63**(3): 451–456.
83. Meiler J, Baker D. (2005) The fumarate sensor DcuS: progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy. *J Magn Reson* **173**(2): 310–316.
84. Alber F, Dokudovskaya S, Veenhoff LM, *et al.* (2007) The molecular architecture of the nuclear pore complex. *Nature* **450**(7170): 695–701.
85. Devos D, Dokudovskaya S, Williams R, *et al.* (2006) Simple fold composition and modular architecture of the nuclear pore complex. *Proc Natl Acad Sci USA* **103**(7): 2172–2177.
86. Andre I, Bradley P, Wang C, Baker D. (2007) Prediction of the structure of symmetrical protein assemblies. *Proc Natl Acad Sci USA* **104**(45): 17656–17661.
87. Diemand AV, Lupas AN. (2006) Modeling AAA+ ring complexes from monomeric structures. *J Struct Biol* **156**(1): 230–243.

88. Moult J. (2005) A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* **15**(3): 285–289.
89. Moult J, Fidelis K, Kryshtafovych A, *et al.* (2007) Critical assessment of methods of protein structure prediction-round VII. *Proteins* **69**(S8): 3–9.
90. Rychlewski L, Fischer D. (2005) LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci* **14**(1): 240–245.
91. Koh IY, Eyrich VA, Marti-Renom MA, *et al.* (2003) EVA: evaluation of protein structure prediction servers. *Nucl Acids Res* **31**(13): 3311–3315.
92. Kryshtafovych A, Fidelis K, Moult J. (2007) Progress from CASP6 to CASP7. *Proteins* **69**(S8): 194–207.
93. Battey JN, Kopp J, Bordoli L, *et al.* (2007) Automated server predictions in CASP7. *Proteins* **69**(S8): 68–82.
94. Kopp J, Bordoli L, Battey JND, *et al.* (2007) Assessment of CASP7 predictions for template-based modeling targets. *Proteins: Struct Funct Bioinform* **69**(S8): 38–56.
95. Hillisch A, Pineda LF, Hilgenfeld R. (2004) Utility of homology models in the drug discovery process. *Drug Discov Today* **9**(15): 659–669.
96. Vangrevelinghe E, Zimmermann K, Schoepfer J, *et al.* (2003) Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J Med Chem* **46**(13): 2656–2662.
97. Feyfant E, Sali A, Fiser A. (2007) Modeling mutations in protein structures. *Protein Sci* **16**(9): 2030–2041.
98. Wattenhofer M, Di Iorio MV, Rabionet R, *et al.* (2002) Mutations in the TMPRSS3 gene are a rare cause of childhood nonsyndromic deafness in Caucasian patients. *J Mol Med* **80**(2): 124–131.
99. Raimondo D, Giorgetti A, Giorgetti A, *et al.* (2007) Automatic procedure for using models of proteins in molecular replacement. *Proteins* **66**(3): 689–696.
100. Poole AM, Ranganathan R. (2006) Knowledge-based potentials in protein design. *Curr Opin Struct Biol* **16**(4): 508–513.
101. Thorsteinsdottir HB, Schwede T, Zoete V, Meuwly M. (2006) How inaccuracies in protein structure models affect estimates of protein-ligand interactions: computational analysis of HIV-1 protease inhibitor binding. *Proteins* **65**(2): 407–423.
102. Herzyk P, Hubbard RE. (1995) Automated method for modeling seven-helix transmembrane receptors from experimental data. *Biophys J* **69**(6): 2419–2442.
103. Peitsch MC, Herzyk P, Wells TN, Hubbard RE. (1996) Automated modelling of the transmembrane region of G-protein coupled receptor by Swiss-model. *Receptors Channels* **4**(3): 161–164.

104. Dahl SG, Edvardsen O, Sylte I. (1991) Molecular dynamics of dopamine at the D2 receptor. *Proc Natl Acad Sci USA* **88**(18): 8111–8115.
105. Henderson R, Schertler GF. (1990) The structure of bacteriorhodopsin and its relevance to the visual opsins and other seven-helix G-protein coupled receptors. *Philos Trans Roy Soc London B Biol Sci* **326**(1236): 379–389.
106. Pebay-Peyroula E, Rummel G, Rosenbusch JP, Landau EM. (1997) X-ray structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipidic cubic phases. *Science* **277**(5332): 1676–1681.
107. Palczewski K, Kumasaka T, Hori T, *et al.* (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* **289**(5480): 739–745.
108. Ballesteros J, Palczewski K. (2001) G protein-coupled receptor drug discovery: implications from the crystal structure of rhodopsin. *Curr Opin Drug Discov Devel* **4**(5): 561–574.
109. Oliveira L, Hulsen T, Lutje Hulsik D, *et al.* (2004) Heavier-than-air flying machines are impossible. *FEBS Lett* **564**(3): 269–273.
110. Cherezov V, Rosenbaum DM, Hanson MA, *et al.* (2007) High-resolution crystal structure of an engineered human {beta}2-adrenergic G protein coupled receptor. *Science* **318**: 1258–1265.
111. Rosenbaum DM, Cherezov V, Hanson MA, *et al.* (2007) GPCR engineering yields high-resolution structural insights into {beta}2 adrenergic receptor function. *Science* **318**: 1266–1273.
112. Peitsch MC. (1995) Protein modelling by E-Mail. *BioTechnology* **13**: 658–660.
113. Guex N, Peitsch MC. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**(15): 2714–2723.
114. Schwede T, Kopp J, Guex N, Peitsch MC. (2003) SWISS-MODEL: An automated protein homology-modeling server. *Nucl Acids Res* **31**(13): 3381–3385.
115. Eswar N, John B, Mirkovic N, *et al.* (2003) Tools for comparative protein structure modeling and analysis. *Nucl Acids Res* **31**(13): 3375–3380.
116. Peitsch MC, Tschopp J. (1995) Comparative molecular modelling of the Fas-ligand and other members of the TNF family. *Mol Immunol* **32**(10): 761–772.
117. Peitsch MC, Wilkins MR, Tonella L, *et al.* (1997) Large-scale protein modelling and integration with the SWISS-PROT and SWISS-2DPAGE databases: the example of *Escherichia coli*. *Electrophoresis* **18**(3–4): 498–501.
118. Sanchez R, Sali A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc Natl Acad Sci USA* **95**(23): 13597–13602.
119. Sanchez R, Pieper U, Mirkovic N, *et al.* (2000) MODBASE, a database of annotated comparative protein structure models. *Nucl Acids Res* **28**(1): 250–253.

120. Kopp J, Schwede T. (2006) The SWISS-MODEL Repository: new features and functionalities. *Nucl Acids Res* **34**(Database Issue): D315–318.
121. Wallner B, Larsson P, Elofsson A. (2007) Pcons.net: protein structure prediction meta server. *Nucl Acids Res* **35**(Web Server Issue): W369–W374.
122. Fox JA, McMillan S, Ouellette BF. (2006) A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucl Acids Res* **34**(Web Server Issue): W3–W5.
123. Berman HM, Burley SK, Chiu W, J *et al.* (2006) Outcome of a workshop on archiving structural models of biological macromolecules. *Structure* **14**(8): 1211–1217.
124. Kopp J, Schwede T. (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucl Acids Res* **32**(Database Issue): D230–D234.
125. Pieper U, Eswar N, Davis FP, *et al.* (2006) MODBASE: A database of annotated comparative protein structure models and associated resources. *Nucl Acids Res* **34**(Database Issue): D291–D295.
126. Sanchez R, Sali A. (1999) ModBase: A database of comparative protein structure models. *Bioinformatics* **15**(12): 1060–1061.
127. Rawlings ND, Morton FR, Barrett AJ. (2006) MEROPS: the peptidase database. *Nucl Acids Res* **34** (Database Issue): D270–D272.
128. Krieger E, Koraimann G, Vriend G. (2002) Increasing the precision of comparative models with YASARA NOVA — a self-parameterizing force field. *Proteins* **47**(3): 393–402.
129. Misura KM, Baker D. (2005) Progress and challenges in high-resolution refinement of protein structure models. *Proteins* **59**(1): 15–29.
130. Baumeister W. (2004) Mapping molecular landscapes inside cells. *Biol Chem* **385**(10): 865–872.
131. Betts MJ, Russell RB. (2007) The hard cell: from proteomics to a whole cell model. *FEBS Lett* **581**(15): 2870–2876.
132. Alber F, Dokudovskaya S, Veenhoff LM, *et al.* (2007) Determining the architectures of macromolecular assemblies. *Nature* **450**(7170): 683–694.
133. Kim DE, Chivian D, Baker D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucl Acids Res* **32**(Web Server Issue): W526–W531.
134. Soding J, Biegert A, Lupas AN. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucl Acids Res* **33** (Web Server Issue): W244–W248.
135. Vriend G. (1990) WHAT IF: a molecular modeling and drug design program. *J Mol Graph* **8**(1): 52–56, 29.