HHS Public Access

Author manuscript

Curr Top Med Chem. Author manuscript; available in PMC 2018 May 10.

Published in final edited form as:

Curr Top Med Chem. 2010 ; 10(1): 84–94.

Template-based Protein Modeling: Recent Methodological Advances

Pankaj R. Daga[§,1], **Ronak Y. Patel**[§,1], and **Robert J. Doerksen**[*,1,2]

[1]Department of Medicinal Chemistry, School of Pharmacy, University of Mississippi, University, MS, 38677-1848

[2]Research Institute of Pharmaceutical Sciences, University of Mississippi

Abstract

Protein modeling has been a very challenging problem in drug discovery and computational biology. The latest advances and progress in computational power have helped to solve this problem to a considerable extent; however, predicting accurate three-dimensional structure of proteins has always been and remains a complicated assignment. Of the two common methods of protein structure prediction, template-based modeling has become more popular than *ab initio* modeling. In this review, we summarize the developments in methodology and of understanding for comparative protein modeling during the last three years, including for homologue search, fold recognition, secondary structure prediction, model building, loop building, side-chain prediction and model quality assessment.

Keywords

Comparative modeling; Fold recognition; Homology modeling; Loop prediction; Quality assessments; Side-chain prediction

Introduction

A protein folds in microseconds to a single well-defined three-dimensional (3D) structure, but how can we reliably predict the 3D structure from the primary sequence? It is important to be able to do so, since knowledge of the 3D structure of proteins aids in elucidating their properties, behavior and almost all biological phenomena mediated by proteins, including protein-ligand and protein-protein interactions. Drug discovery and protein design also benefit from knowledge of 3D structure. Computational techniques for protein structure prediction are necessary because it has thus far proven impossible for experimental structure determination to keep pace with the increasing number of reported protein sequences. Experimental techniques have advanced and there are now more than 53,000 experimentally solved 3D structures in the Protein Data Bank (PDB), but there are far more protein sequences being reported, with >400,000 manually annotated and reviewed in Swiss-Prot

[*]CORRESPONDING AUTHOR: Dr. Robert J. Doerksen, University of Mississippi, Department of Medicinal Chemistry, 421 Faser Hall, University, MS 38677, Office: 662-915-5880, Fax: 662-915-5638, rjd@olemiss.edu.

[§]These authors contributed equally to this work.

and >7,500,000 automatically annotated in TrEMBL.[1] 880 genome projects have been completed and more than twice that many are at various stages of completion. The gap between sequence and structure can be bridged by structure prediction methods.[2, 3] This objective is realistic considering that sequences with 50% identity can often be modeled within experimental accuracy.[4]

Structure prediction methods are broadly classified into homology modeling (HM), also called template-based modeling (TBM) or comparative modeling (CM), and free modeling (FM).[5] A basic assumption is that proteins with similar sequence fold into similar 3D structures. In HM, the 3D structure of the protein is built commencing from structural information of evolutionarily-related sequence(s), whereas the more general names TBM or CM denote that a template protein is used but that the template is not necessarily of related history or function to the target. FM does not depend on *a priori* structural information and hence often yields less reliable models than TBM; this review focuses on TBM, so topics related to FM are not included. TBM involves several steps: identification of homologues (templates), alignment of target to template, structure building, refinement and validation. Prior excellent reviews have been published on this topic.[2–11] Extensive reviews about the available programs for homology modeling have been reported.[7, 12, 13] This review summarizes key recent methodological advances from the last three years for the various steps of TBM. Some of the methods have been listed in Table 1, which gives a convenient summary of relevant web servers and/or programs.

Critical assessment of structure prediction (CASP)

CASP is a series of protein structure prediction community experiments conducted every two years since 1994. Different groups developing structure prediction methods submit the results of their predictions for competition targets whose experimental structures are available but not yet disclosed. The development of CASP over the years is an index of the development of template-based structure prediction methodologies and capabilities.[10, 14] For CASP7 relative to CASP6, comparison of closeness of models to experimental structures showed that there have been improvements, especially for medium or high difficulty targets.[4] When the sequence identity between target and available templates was high, the best predicted models were within experimental accuracy. Performance of CASP8 is somewhat similar to CASP7 in this regard. Improvements were observed throughout the CASP experiments for alignment accuracy and modeling of regions not available from the template.[4, 15, 16] Particularly notable have been improvements in the performance of fully automated servers. Though the best predictions were those that used human expertise, for many of the targets fully automated servers performed well also.[4, 15] Out of the best six predicted structures by humans or automated servers in CASP7 and CASP8, ~29% were from the servers. This is a clear improvement over CASP5 and CASP6, in which the number was ~15%. In addition to that, for 90% of the CASP8 targets at least one of the top six predictions was from an automated server, which is a significant improvement over previous CASPs.[4, 15, 16] Such success is welcome especially for large scale modeling approaches, in which reliance on human expertise can be prohibitively expensive.

Homologue search and fold recognition

The first step towards prediction of protein structure using homology modeling is identification of a template, a homologous protein with known 3D structure, by searching an available database of sequences. Detection of homologues is generally accompanied by sequence alignment, which can be used subsequently in 3D structure construction, though the alignment may require editing. Homologue search is generally based on sequence similarity between template and target or on features that describe the physicochemical nature of amino-acids such as secondary structure and solvent accessibility. A sequence based homologue search can be performed using a single target sequence (BLAST[17], FASTA[18]) or using a profile generated from multiple sequences (PSI-BLAST[19], HMMER[20], the latter using hidden Markov models (HMM)). Use of profiles and discriminative algorithms such as support vector machines (SVM) has greatly improved the capability for remote homologue detection. Fold recognition methods, generally suggested for automatic prediction of protein structure, can also be used as a tool for identifying remote homologues.[7] In this review article, we have focused on methods which search for homologues in the PDB and provide pair-wise sequence alignment of target and template, and often a built model using that alignment (fold recognition).

In one important development, using the meta-server format that relies on the consensus of outputs of different servers for fold prediction, an ensemble of 31 independent algorithmic variants for remote homology detection were generated and used for fold recognition in Phyre software.[21] The alignments generated were used for building 3D structures followed by selection of the best model using 3D-Colony, which uses both a measure of fold recognition assignment confidence as well as structural similarity clustering. PDBalert[22] performs an HMM HHpred[23] search for a query sequence against PDB. If HHpred does not detect homologues, PDBalert will continue the search every week and then notify the user as soon as a homologous protein structure is made available in PDB. The Zhou group has developed a series of methods based on weighted matching of sequence and a structure based profile. Recently, the use of solvent accessibility, residue depth profiles, and torsion angles profiles led to the development of SP4[24] and SP5.[25] Incorporation of these parameters in the parent method has been shown to improve prediction of remote homologues and alignment accuracy. In a somewhat slower approach, FoldPro derives a set of pair-wise features comprised of different similarity scores.[26] The structural relevance of target-template pair-wise alignment is then checked using a supervised classification method based on feature vectors, followed by selection of the best templates from the database. COMPASS, which identifies the fold of a protein using profile-profile alignment of target and template, is now available as a web server and provides improved selectivity and sensitivity of homologue detection.[27]

Several papers have reported new methods for detection of remote homology that are primarily used for function prediction, identification of protein families or classification of protein sequences into families.[28–39] The methods can be used for identifying the family, super-family or fold of target proteins which might be a remote homologue, in cases when conventional methods fail to identify a suitable template. The target protein can then be

aligned to the template using sequence-sequence, sequence-profile or profile-profile alignments.

Earlier literature on sequence comparison and alignment has been excellently reviewed.[6, 40] As an attempt to obtain better alignment of target-template, the use of the generalized Viterbi algorithm with a hidden Markov model has led to the development of HMM-Kalign, a part of HMMER.[41] This algorithm explores suboptimal alignments, unlike standard HMM which always selects the sequence with the best score. Use of HMM-Kalign to generate an alignment used for homology modeling of oxidized bacteriophage T4 glutaredoxin led to a lower RMSD relative to the available crystal structure compared to optimized sequence alignment generated using HMM.[41] The performance of profile-profile alignment for fold recognition and remote homologue detection compared to simple profile-profile alignment (PSI-BLAST and HHsearch) was greatly improved using nonnegative matrix factorization.[42] Use of structure alignment instead of sequence alignment for profile generation improved performance in low identity regions.[43] There have also been efforts to construct new substitution matrices for pair-wise or multiple sequence alignment, which can enhance alignment quality compared to traditional PAM or BLOSUM matrices.[44]

Secondary structure prediction

Secondary structure prediction (SSP) is important for the generation of optimal alignment and selection of suitable template(s) and, subsequently, for successful structure prediction. [45–47] The methods for SSP assume that multiple homologous sequences have a similar structure.[48] Most of the recently developed methods are based on the use of multiple sequences[49–56] with a few using a single sequence.[57, 58]

Recent developments of SSP methods have led to prediction accuracy up to 80%, approaching a reported theoretical limit of 88% from available 3D structure.[59] Improvement of BSPSS into IPSSP using new learning model and training methods[57] led to increased three state (helices, sheets and loops) prediction accuracy on a test dataset. MUPRED is a combination of fuzzy *k*-nearest neighbor (FKNN) and profile based methods [prediction accuracy (PA): 79.2–81.1%].[53] MUPRED can be used for query sequences having many, few or no homologues. Profile based methods are used when many homologues are available; FKNN dominates the prediction when few or no homologues are present. In a similar approach, consensus data mining (CDM) has been developed, which combines fragment data mining (FDM) and GOR V (Garnier-Osguthorpe-Robson information theory/Bayesian method), for SSP. If the identity at a particular position is higher than the overall sequence identity, then the prediction by FDM is used as the final prediction; if lower, the prediction by GOR V is used (PA: 68–93%).[47, 51] The use of PSI-BLAST and HMMER profiles instead of previously used frequency profiles in the revised Jpred algorithm led to improved accuracy (PA: 81.5%).[52] PROTEUS is a combination of methods in which the first step is secondary structure prediction for regions which can be mapped onto known 3D structures. In the second step, predictions of three algorithms (Jnet, PSIPRED and TRANSSEC) are combined in a neural network to predict the secondary structure. A final prediction is made by combining both the steps (PA: 81.3%).[50]

Porter_H,[49] an improvement of Porter[60], is a method based on *ab initio* secondary structure and homology based prediction. When sequence similarity between query and structure homologue is > 30%, this modification improved the prediction (PA: 90%). Extreme learning methods have been shown to perform as well as existing methods (PA: 71.2%).[58] A few more methods based on a genetic algorithm (PA: 75.1%) [61], neural networks (PA: 78.1%)[56] or SVM (PA: 82.2%)[62] were proposed with moderate to high accuracy. A two stage Multi-class SVM was also fairly accurate (PA: 77.0% to 79.5%).[63]

Model Building

After sequence alignment between the query sequence and template sequence, the next step is model building. The four principal methods for model construction include spatial restraint method (SSR),[64] segment matching method (SMM),[65] multiple template method (MTM) [66, 67] and artificial evolution (AE).[68]

SSR assumes that several geometrical features such as distances and angles are conserved in homologous proteins, when comparing equivalent positions. SSR methodology involves two main steps, extraction of spatial restraints based on alignment and construction of the target 3D-model by fulfilling the spatial restraints.[64] MODELLER, currently on version 9v6, uses SSR and is one of the most frequently used homology modeling programs today.

SMM divides the target into a series of short segments, each matched to its own template fitted from the PDB. Sequence alignment is done over segments rather than over the entire protein. Different steps in this method include constructing the segment database, model construction via iterative randomization to get an average model and minimization to get the final model. Recently, Larsson et al proposed Pfrag, an extension to the SMM program SegMod/ENCAD which can use multiple templates.[69]

In MTM, several solved protein 3D structures are used to build the target protein model. The multiple templates are aligned with each other based on sequences and structures. The target is optimally aligned with the multiple templates. Structural alignment of the homologous proteins reveals structural elements that are conserved in all the templates, which are mainly composed of secondary structural elements. Structurally variable regions (loops) are present between the conserved regions. The loops are usually exposed at the surface of the proteins. This method has been implemented in several packages such as 3D-JIGSAW,[70] SWISS-MODEL[71] and MOE.[72]

In AE, the alignment of template and target sequences is carried out using the concepts of evolution: mutations, insertions and deletions. The target protein model is built by editing the template structure based on the alignment. The iterative process starts with simple mutations of the aligned residues (surface residues followed by buried residues), the operation which least changes the energy, and subsequent minimization. Mutations are followed by deletions and then insertions. Deletion is prioritized over insertion since it is more easily predicted. For each step, an operation is considered successful if it does not cause a significant energy penalty; the whole process is repeated until the final model is

obtained.[68] NEST, the core program in the JACKAL modeling package, uses the AE method.

In recent benchmarking studies, the performance of six homology modeling programs was compared. MODELLER, NEST and SegMod/ENCAD were found to perform the best.[12] According to results in CASP7, some automated web-servers have been very successful in accurate prediction of the protein target structures. In particular, I-TASSER, ROBETTA, Pmodeller-6 implemented in the consensus server Pcons, and HHPred3 performed very well. I-TASSER searches the whole PDB library to find appropriate protein fragments. Matching aligned fragments are combined to assemble the global structure, while for portions for which no alignment matches are found, the 3D structure is built using *ab initio* simulations. Final refinement of the model consists of lowest energy conformational search.[73, 74] Model building in Pcons is carried out using Pfrag,[69] a modified SegMod homology modeling program, and final refinement is performed using the ENCAD force field.[75] Model prediction by ROBETTA makes use of extensive and computationally expensive conformational sampling and all-atom energy refinement.[76]

In recent developments, a modification to TASSER, TASSER-Lite [77] was proposed for faster protein model construction compared to the original method. It is appropriate for modeling of proteins for which a highly homologous template is available, since extensive conformational searching is avoided.[77] The M4T (Multiple Mapping Method with Multiple Templates) web-based homology modeling server included two major modules, Multiple Templates (MT) and Multiple Mapping Method (MMM).[78] The MT module selects and optimally combines the sequences of multiple template structures while MMM improves alignment accuracy. Final model building in M4T is carried out using MODELLER. The PROTEUS2 web-server combined various tools including transmembrane helix and β-strand prediction, SSP and 3D structure prediction, using machine learning and database comparison techniques.[79]

Loop Modeling

Protein loops connect well-defined secondary structure regions such as α-helices and β-sheets. Loops are comparatively difficult to study by X-ray crystallography and often represent poorly conserved regions in a given family of proteins. Loops play a wide variety of roles related to protein function, ligand binding sites or active sites[80] and thus can play a critical role in structure-based design. Loop model building in homology models, because of structural inaccuracies in the models, is more difficult than loop reconstruction methods tested on crystal structures.[81] Loop modeling methods can be classified into two major approaches: (i) Knowledge based and (ii) energy based. A few methods have also been reported which combine the two approaches. Recent reports have reviewed these methods comprehensively.[82, 83] Knowledge based methods find from a database of structures a loop segment that fits between the two stem regions of the loop. These methods are mainly limited by the availability of relevant loop structures from known protein structures.[84] Recently, a classification database of structural motifs, ArchDB, was developed[85] and evaluated using two different sequence profiles, and a hidden Markov model (HMM) profile was found to produce encouraging results.[80] A hierarchical and multidimensional database

has been reported that classified more than 100,000 loop fragments. The loop length, types of bracing structures, and geometric restraints of stems helped in loop selection and a Z-score provided the final ranking of the loops.[86] The use of an artificial neural network to evaluate sequence-template alignments of loops has been proposed and gave accurate predictions.[84] Another method used Monte Carlo simulation of the loop, commencing loop prediction with fragment databases ranked using the DFIRE potential.[87]

Energy based methods use an *ab initio* energy function for conformational search of the loops and to judge their quality. A recent report gives a overview of existing *ab initio* methods.[83] *Ab initio* loop modeling has proven to be very accurate for prediction of loops as long as seven residues,[88] but is less reliable for longer loops. Loop conformational search can be carried out using numerous available tools such as local move Monte Carlo (LMMC),[89] torsion angle conformational search,[90] the Direct Tweak algorithm in LoopBuilder,[91] replica exchange[92] or a dihedral angle-based buildup procedure in hierarchical loop prediction (HLP).[88] The generated conformers are scored using force field or other physics-based energy calculations[93] usually including solvation effects. The systematic and efficient sampling strategy in a newly developed protocol, LOOPER, searched for loop conformers with optimal interactions of the loop backbone with the rest of the protein atoms. Final ranking in LOOPER is carried out using a CHARMm energy scoring function with a generalized Born solvation term.[94] A few modifications of the Protein Local Optimization Program (PLOP) have been reported to improve loop prediction by use of a novel solvent model[90] or energy model.[95]

Side-chain Modeling

Side-chains play a major role in drug design, such as in ligand docking,[96] in protein modeling for loop building[88] and in general for prediction of protein structures.[97] Most of the methods for side-chain prediction use rotamer libraries which are constructed reliably using statistical knowledge of protein 3D structures. A number of methods have been developed for rotamer-based side-chain modeling, and recent developments include enhanced sampling schemes and the use of modified energy and/or scoring functions. A few methods include a combination of the approaches.[98, 99] Recent rotamer-library independent methods include the Grow-to-Fit molecular dynamics method (G2FMD)[100] and statistical machine learning methods.[101]

Various rotamer-based methods have been reviewed[102] and will not be mentioned here. The recently proposed method IRECS selects more than one rotamer for the side-chain in order to have a representation of the conformational space flexibility of the side-chain. IRECS ranking is provided by a knowledge based statistical potential, ROTA.[103] ChiRotor, for rapid prediction of side-chains, uses a limited sampling procedure in combination with energy minimization.[104] A combination of residue reduction and rotamer reduction was developed for efficient and accurate side-chain prediction.[105]

Modifications to the ROSETTA energy functions, with softer van der Waals terms, and extended rotamer libraries have been implemented to improve side-chain modeling.[106] Based on Tree Reweighted Belief Propagation, a novel search method and novel energy

(modified ROSETTA) function were proposed to predict global minima more reliably.[98] Variations in the internal dielectric constant of a protein and the use of and quality of solvent model employed have been shown to play significant roles in the prediction of side-chains.[95, 107] Another novel method, OPUS-Rota, combines the newly introduced OPUS-PSP potential with heat bath Monte Carlo conformational search.[99]

Quality Assessment (QA)

Methods of protein model QA have recently been reviewed.[11] A comparative model generated through one or more steps described above may have incorrect geometry and energy. Some common types of errors that can occur during model building are use of incorrect alignment or template, errors in portions of the target structure built without template, distortion in correctly aligned regions, or errors in side-chain packing.[46] Since protein structure prediction methodology has been extensively automated, there is more need for software/servers which can select the stereochemically or energetically best models from among decoys built using the same or different templates. Methods of protein QA can be either statistical or physico-chemical, and could be based on alignment to a single template or multiple templates or on metaserver results. The QA method either gives a local score as a function of residue or residue window[108–112] or a global score[113–116] which may be based on single or multiple assessment criteria.

The Undertaker program, using a genetic algorithm, generates 3D structures using an alignment suggested by the Sequence Alignment and Modeling system (SAM) HMM package.[117] The best structures are selected by the undertaker cost function, comprised of 73 individual cost functions. In another report addressing improvement of the undertaker scoring function, weighted distance constraints generated from alignment to different templates were used for model quality assessment.[118] The globularity index, a combined score including hydrogen bonding information, solvent accessible surface area, voids and the number of water molecules within 5 Å of the protein, has been used for evaluation of the quality of protein models.[119] The ModFOLD[108] server combines ModSSEA[120], MODCHECK[121] and ProQ[122] scores with secondary structure information. ModFOLDclust performs clustering of multiple models and calculates per residue scores which depict the local quality of a cluster. A reduced representation of statistical potential such as a C_β potential (in which side-chain atoms beyond C_β itself are ignored) is simpler and computationally less intensive than an all-atom potential representation. Information about backbone geometry and primary sequence separation was incorporated to obtain an improved C_β potential.[123] The improved reduced potential has been shown to outperform the DOPE all-atom potential[124] for identification of native structure from among decoys. Fifteen parameters based on energy, secondary structure, solvent accessible surface area and hydrophobic contact have been implemented using a neural network in the Artificial Intelligence Decoys Evaluator (AIDE).[111] AIDE showed similar or better performance to ProQ[122] and Victor[125] on test datasets. The local quality of a structure can be quantified using ProQres, which relies on 3D information, or ProQprof, which utilizes a model generated from sequence alignment.[110] ProQres quantifies structural qualities such as secondary structure, solvent accessibility, and atom-atom and residue-residue contacts to have a measure of local quality. ProQprof uses profiles both for target and template. A sum

of the two scores has also been proposed as ProQlocal. ModelEvaluator quantifies the absolute quality of a protein model using support vector regression (SVR),[126] and was trained using only structural characteristics such as secondary structure, contact map, relative solvent accessibility and beta sheet structure. The score is suitable for comparing the quality of 3D structures of different proteins. The C_α potential and fragment quality comparison have been used to select the best model from a set of structures generated by different methods (using TASSER-QA). This approach performed as well as other QA methods for assessing the CASP7 test sets, and exceeded their performance for medium and hard targets.[127]

QMEAN is another composite scoring function composed of five protein geometry structural descriptors, including a new torsional angle potential treating groups of three sequential amino-acids.[113] A weighted sum of six quality assessment scores has been implemented in SVMod.[114] The best support vector machine (SVM) score outperformed other tested physical, statistical and machine learned individual scores for selection of the best model (closest to the native structure) from among decoys. In an effort to measure the local quality of modeled structures, SVM has been trained using the DFIRE contact and torsional statistical potential[128] along with attributes describing information about the local environment, resulting in better scoring than other local QA methods.[129] FragQA uses the C_α RMSD between template and modeled target structure to predict the local quality of modeled structures in regions where there are no alignment gaps between target and template.[109] Comparison with ProQres[12] showed equal performance of FragQA for the test set used.

SuboPtimal Alignment Diversity (SPAD) quantifies the error in modeled structures based on alignment stability, by estimating how well suboptimal alignments converge to the optimal alignment.[130] Since this method calculates the SPAD score using the target/template alignment, not the 3D structures, it can be used to estimate probable errors during early stages of homology modeling. Fams-ace, a collection of meta server based tools, has been improved as Fams-ace (improved).[131] The major improvement was incorporation of a new method for final model selection, CIRCLE, which uses a knowledge-based potential for side-chain packing. Meta-MQAP uses a multivariate regression model of eight model quality assessment programs (MQAP) with correction of trivial errors caused by any of the programs.[132] Assessment of model quality using Meta-MQAP showed a very good correlation with actual deviation from native structure. SELECTpro, a structure based evaluation method, combines several physical, statistical and predicted structural terms to address protein quality problem.[116]

A few of the above-mentioned methods have been demonstrated to perform equally well or better than the existing methods when tested on CASP results. For example, the globularity index could differentiate well between the good and bad models predicted in CASP6.[119] Similarly, the newly developed methods ModelEvaluator,[129] undertaker cost function,[133] SVR-method,[126] QMEAN,[113] and SELECTpro[116] performed well on the CASP7 dataset and were found to be very effective in selecting models close to the native structures.

Template Based Modeling in Drug Design

TBM has been widely used in the process of drug design and discovery. A recent review has discussed the current trends and applications of homology modeling in the drug discovery process.[134] Various recent applications of homology modeling in the drug discovery process include lead identification,[135, 136] lead optimization, understanding of selectivity,[137] explanation of resistance development,[138] binding site analysis and mutation studies.[139] An example is for the histidine kinase (HK) VicK protein, which is essential for bacterial growth in S. pneumoniae. Li et al. built a comparative model of the HK VicK protein (33% identity with template), which was further successfully used for structure-based virtual screening to identify novel potential HK inhibitors with antibacterial activity.[135] Sharon and Chu have carried out active site analysis to understand the molecular basis of drug resistance using a hepatitis B virus (HBV) DNA polymerase model (built using HIV-1 reverse transcriptase with <20% identity). Certain amino acid mutations were found to be responsible for the resistance development against marketed anti-HBV drugs. This study, using wild-type as well as mutant HBV polymerases, suggested a significant correlation between the fold resistances and the protein-ligand binding affinity of anti-HBV nucleosides.[138]

Among the applications listed above, the most important is in structure based drug discovery. In order to model ligand binding sites, usually ligand coordinates are manually added or automatically docked to the best model generated and next refined by local energy minimization. The generated model can then be used for further applications.[140–142] Some alternate approaches to sample different conformations of binding sites (or of the entire protein) are molecular dynamics simulation (MD),[143, 144] normal mode analysis (NMA)[145] and generation of a series of models.[146, 147] Homology models generated in this way are often referred to as ligand steered homology models[146] or ligand-supported homology models.[140] Although the former method (coordinate transfer and local minimization) is simple and only a single model is generated, the latter set of methods (MD, NMA, multiple models) generate several homology model (conformations), which can make further studies such as docking or virtual screening time-consuming. Application of these methods to virtual screening of millions of compounds is impossible at present but they can be used for small library screening[143, 146] and binding pose prediction for one or a few ligands.

Conclusion

The exponentially-increasing difference between number of protein sequences and available experimental 3D structures makes it obligatory to rely on protein modeling methods to build 3D protein models. Advances in computational power and innovation have led to the development of novel and accurate methods for the 3D modeling of proteins. Some of the new methods have been proven to be accurate and rapid. However, the capability to be able to build a protein model very close to the native structure of the protein reliably is a challenging assignment and is still under development. Many methods are available on the internet, as listed in Table 1. Accurate prediction of protein modeling will certainly assist in understanding the mechanism of action of proteins and will aid in drug design to devise

better ligands for the protein. Considering all the developments in the field, the task seems to be achievable in the near future.

Acknowledgments

Funding from University of Mississippi; National Center for Zoonotic, Vector-borne, and Enteric Diseases (CK) of the Centers for Disease Control and Prevention (CDC) (U50/CCU423310); National Science Foundation EPS-0556308; and from the National Institute of Health's National Center for Research Resources (P20 RR021929 and C06 RR-14503-01) are greatly appreciated. PRD is a CoBRE CORE-NPN fellow.

List of Abbreviations

AIDE	Artificial Intelligence Decoys Evaluator
BLAST	Basic Local Alignment Search Tool
BLOSUM	BLOck SUbstitution Matrix
BSPSS	Bayesian Segmentation of Protein Secondary Structure
CASP	Critical Assessment of Structure Prediction
CM	Comparative Modeling
COMPASS	Comparison of Multiple Protein Alignments with Assessment of Statistical Significance
DEE	Dead-End-Elimination
FKNN	Fuzzy *k*-nearest neighbor
FM	Free Modeling
G2FMD	Grow-to-Fit Molecular Dynamics Method
GOR	Garnier-Osguthorpe-Robson
HHpred	*pred*iction by HMM-HMM comparison
HMM	Hidden Markov model
IPSSP	Iterative Protein Secondary Structure Parse
IRECS	Iterative REduction of Conformational Space
M4T	Multiple Mapping Method with Multiple Templates
MQAP	Model Quality Assessment Programs
MUPRED	PREDiction software from University of Missouri
PA	Prediction Accuracy
PAM	Point Accepted Mutation
PSI-BLAST	Position Specific Iterative BLAST

QMEAN	Qualitative Model Energy Analysis
RMSD	Root Mean Squared Deviation
SAM	Sequence Alignment and Modeling
SPAD	SuboPtimal Alignment Diversity
SVM	Support Vector Machines
TBM	Template-Based Modeling

References

1. Jain E, Bairoch A, Duvaud S, Phan I, Redasch N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E. Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics. 2009; 10:136. [PubMed: 19426475]
2. Zhexin X. Advances in Homology Protein Structure Modeling. Curr Prot Pept Sci. 2006; 7(3):217–227.
3. Ginalski K. Comparative modeling for protein structure prediction. Curr Opin Struct Biol. 2006; 16(2):172–177. [PubMed: 16510277]
4. Kryshtafovych A, Fidelis K, Moult J. Progress from CASP6 to CASP7. Proteins. 2007; 69(Suppl 8): 194–207. [PubMed: 17918728]
5. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol. 2008; 18(3):342–348. [PubMed: 18436442]
6. Dunbrack JRL. Sequence comparison and protein structure prediction. Curr Opin Struct Biol. 2006; 16(3):374–384. [PubMed: 16713709]
7. Esposito EX, Tobi D, Madura JD. Comparative protein modeling. Rev Comput Chem. 2006; 22:57–167.
8. Fischer D. Servers for protein structure prediction. Curr Opin Struct Biol. 2006; 16(2):178–182. [PubMed: 16546376]
9. Floudas CA. Computational methods in protein structure prediction. Biotechnol Bioeng. 2007; 97(2):207–213. [PubMed: 17455371]
10. Guo JT, Ellrott K, Xu Y. A historical perspective of template-based protein structure prediction. Methods Mol Biol. 2008; 413:3–42. [PubMed: 18075160]
11. Kryshtafovych A, Fidelis K. Protein structure prediction and model quality assessment. Drug Discov Today. 2009; 14(7–8):386–393. [PubMed: 19100336]
12. Wallner B, Elofsson A. All are not equal: A benchmark of different homology modeling programs. Protein Sci. 2005; 14(5):1315–1327. [PubMed: 15840834]
13. Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L. Sequence based residue depth prediction using evolutionary information and predicted secondary structure. BMC Bioinformatics. 2008; 9:388. [PubMed: 18803867]
14. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol. 2005; 15(3):285–289. [PubMed: 15939584]
15. Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T. Automated server predictions in CASP7. Proteins. 2007; 69(Suppl 8):68–82. [PubMed: 17894354]
16. Kryshtafovych A, Fidelis K, Moult J. CASP8 results in context of previous experiments. Proteins. 2009
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Bio. 1990; 215(3):403–410. [PubMed: 2231712]
18. Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics. 1991; 11(3):635–650. [PubMed: 1774068]

19. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. Nucleic Acids Res. 1997; 25(17):3389–3402. [PubMed: 9254694]
20. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998; 14(9):755–763. [PubMed: 9918945]
21. Bennett-Lovsey RM, Herbert AD, Sternberg MJ, Kelley LA. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins. 2008; 70(3):611–625. [PubMed: 17876813]
22. Agarwal V, Remmert M, Biegert A, Söding J. PDBalert: Automatic, Recurrent Remote Homology Tracking and Protein Structure Prediction. BMC Struct Biol. 2008; 8:51. [PubMed: 19025670]
23. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005; 21(7):951–960. [PubMed: 15531603]
24. Liu S, Zhang C, Liang S, Zhou Y. Fold recognition by concurrent use of solvent accessibility and residue depth. Proteins. 2007; 68(3):636–645. [PubMed: 17510969]
25. Zhang W, Liu S, Zhou Y. SP5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model. PLoS ONE. 2008; 3(6):e2325. [PubMed: 18523556]
26. Cheng J, Baldi P. A machine learning information retrieval approach to protein fold recognition. Bioinformatics. 2006; 22(12):1456–1463. [PubMed: 16547073]
27. Sadreyev RI, Tang M, Kim BH, Grishin NV. COMPASS server for remote homology inference. Nucleic Acids Res. 2007; 35:W653–658. Web Server issue. [PubMed: 17517780]
28. Dong QW, Wang XL, Lin L. Application of latent semantic analysis to protein remote homology detection. Bioinformatics. 2006; 22(3):285–290. [PubMed: 16317074]
29. Hådstad T, Hestnes AJH, Saetrom P. Motif kernel generated by genetic programming improves remote homology and fold detection. BMC Bioinformatics. 2007; 8:23. [PubMed: 17254344]
30. Casbon JA, Saqi MA. On single and multiple models of protein families for the detection of remote sequence relationships. BMC Bioinformatics. 2006; 7:48. [PubMed: 16448555]
31. Hochreiter S, Heusel M, Obermayer K. Fast model-based protein homology detection without alignment. Bioinformatics. 2007; 23(14):1728–1736. [PubMed: 17488755]
32. Lingner T, Meinicke P. Word correlation matrices for protein sequence analysis and remote homology detection. BMC Bioinformatics. 2008; 9:259. [PubMed: 18522726]
33. Liu B, Wang X, Lin L, Dong Q. A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis. BMC Bioinformatics. 2008; 9:510. [PubMed: 19046430]
34. Söding J, Remmert M, Biegert A, Lupas AN. HHsenser: exhaustive transitive profile search using HMM-HMM comparison. Nucleic Acids Res. 2006; 34:W374–378. Web Server issue. [PubMed: 16845029]
35. Tyagi M, de Brevern AG, Srinivasan N, Offmann B. Protein structure mining using a structural alphabet. Proteins. 2008; 71(2):920–937. [PubMed: 18004784]
36. Melvin I, Ie E, Kuang R, Weston J, Stafford WN, Leslie C. SVM-Fold: a tool for discriminative multi-class protein fold and superfamily recognition. BMC Bioinformatics. 2007; 8(Suppl 4):S2.
37. Damoulas T, Girolami MA. Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. Bioinformatics. 2008; 24(10):1264–1270. [PubMed: 18378524]
38. Oğul H, Mumcuoğlu EU. SVM-based detection of distant protein structural relationships using pairwise probabilistic suffix trees. Comput Biol Chem. 2006; 30(4):292–299. [PubMed: 16880118]
39. Bhadra R, Sandhya S, Abhinandan KR, Chakrabarti S, Sowdhamini R, Srinivasan N. Cascade PSI-BLAST web server: a remote homology search tool for relating protein domains. Nucleic Acids Res. 2006; 34:W143–146. Web Server issue. [PubMed: 16844978]
40. Fariselli P, Rossi I, Capriotti E, Casadio R. The WWWH of remote homolog detection: the state of the art. Brief Bioinform. 2007; 8(2):78–87. [PubMed: 17003074]
41. Becker E, Cotillard A, Meyer V, Madaoui H, Guérois R. HMM-Kalign: a tool for generating sub-optimal HMM alignments. Bioinformatics. 2007; 23(22):3095–3097. [PubMed: 17921492]

42. Jung I, Lee J, Lee SY, Kim D. Application of nonnegative matrix factorization to improve profile-profile alignment features for fold recognition and remote homolog detection. BMC Bioinformatics. 2008; 9:298. [PubMed: 18590572]
43. Bernardes JS, Dávila AM, Costa VS, Zaverucha G. Improving model construction of profile HMMs for remote homology detection through structural alignment. BMC Bioinformatics. 2007; 8:435. [PubMed: 17999748]
44. Tan YH, Huang H, Kihara D. Statistical potential-based amino acid similarity matrices for aligning distantly related protein sequences. Proteins. 2006; 64(3):587–600. [PubMed: 16799934]
45. Aloy P, Mas JM, Marti-Renom MA, Querol E, Aviles FX, Oliva B. Refinement of Modelled Structures by Knowledge-Based Energy Profiles and Secondary Structure Prediction: Application to the Human Procarboxypeptidase A2. J Comput Aided Mol Des. 2000; 14(1):83–92. [PubMed: 10702927]
46. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Ann Rev Biophys Biomol Struct. 2000; 29:291–325. [PubMed: 10940251]
47. Sen TZ, Cheng H, Kloczkowski A, Jernigan RL. A Consensus Data Mining secondary structure prediction by combining GOR V and Fragment Database Mining. Protein Sci. 2006; 15(11):2499–2506. [PubMed: 17001039]
48. Heringa J. Computational methods for protein secondary structure prediction using multiple sequence alignments. Curr Prot Pept Sci. 2000; 1(3):273–301.
49. Pollastri G, Martin AJ, Mooney C, Vullo A. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. BMC Bioinformatics. 2007; 8:201. [PubMed: 17570843]
50. Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS. Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics. 2006; 7:301. [PubMed: 16774686]
51. Cheng H, Sen TZ, Jernigan RL, Kloczkowski A. Consensus Data Mining (CDM) Protein Secondary Structure Prediction Server: combining GOR V and Fragment Database Mining (FDM). Bioinformatics. 2007; 23(19):2628–2630. [PubMed: 17660202]
52. Cole C, Barber JD, Barton GJ. The Jpred 3 secondary structure prediction server. Nucleic Acids Res. 2008; 36:W197–201. Web Server issue. [PubMed: 18463136]
53. Bondugula R, Xu D. MUPRED: a tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. Proteins. 2007; 66(3): 664–670. [PubMed: 17109407]
54. Karypis G. YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction. Proteins. 2006; 64(3):575–586. [PubMed: 16763996]
55. Lin HN, Chang JM, Wu KP, Sung TY, Hsu WL. HYPROSP II–a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. Bioinformatics. 2005; 21(15):3227–3233. [PubMed: 15932901]
56. Yao XQ, Zhu H, She ZS. A dynamic Bayesian network approach to protein secondary structure prediction. BMC Bioinformatics. 2008; 9:49. [PubMed: 18218144]
57. Aydin Z, Altunbasak Y, Borodovsky M. Protein Secondary Structure Prediction for a Single-Sequence Using Hidden Semi-Markov Models. BMC Bioinformatics. 2006; 7:178. [PubMed: 16571137]
58. Wang GR, Zhao Y, Wang D. A protein secondary structure prediction framework based on the Extreme Learning Machine. Neurocomputing. 2008; 72(1–3):262–268.
59. Rost, B. Rising accuracy of protein secondary structure prediction. In: Chasman, D., editor. Protein structure determination, analysis, and modeling for drug discovery. Dekker; New York: 2003. p. 207-249.
60. Pollastri G, McLysaght A. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics. 2005; 21(8):1719–1720. [PubMed: 15585524]
61. Won KJ, Hamelryck T, Prügel-Bennett A, Krogh A. An evolutionary method for learning HMM structure: prediction of protein secondary structure. BMC Bioinformatics. 2007; 8:357. [PubMed: 17888163]

62. Duan M, Huang M, Ma C, Li L, Zhou Y. Position-specific residue preference features around the ends of helices and strands and a novel strategy for the prediction of secondary structures. Protein Sci. 2008; 17(9):1505–1512. [PubMed: 18519808]

63. Nguyen MN, Rajapakse JC. Prediction of Protein Secondary Structure with two-stage multi-class SVMs. Int J Data Min Bioinform. 2007; 1(3):248–269. [PubMed: 18399074]

64. Sali A, Blundell TL. Comparative Protein Modeling by Satisfaction of Spatial Restraints. J Mol Bio. 1993; 234(3):779–815. [PubMed: 8254673]

65. Levitt M. Accurate Modeling of Protein Conformation by Automatic Segment Matching. J Mol Bio. 1992; 226(2):507–533. [PubMed: 1640463]

66. Chothia C, Lesk AM, Levitt M, Amit AG, Mariuzza RA, Phillips SEV, Poljak RJ. The Predicted Structure of Immunoglobulin-D1.3 and its Comparison with the Crystal-Structure. Science. 1986; 233(4765):755–758. [PubMed: 3090684]

67. Blundell TL, Sibanda BL, Sternberg MJE, Thornton JM. Knowledge-Based Prediction of Protein Structures and the Design of Novel Molecules. Nature. 1987; 326(6111):347–352. [PubMed: 3550471]

68. Petrey D, Xiang Z, Tang CL, Xie L, Gimpelev M, Mitros T, Soto CS, Goldsmith-Fischman S, Kernytsky A, Schlessinger A, Koh IY, Alexov E, Honig B. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. Proteins. 2003; 53(Suppl 6):430–435. [PubMed: 14579332]

69. Larsson P, Wallner B, Lindahl E, Elofsson A. Using multiple templates to improve quality of homology models in automated homology modeling. Protein Sci. 2008; 17(6):990–1002. [PubMed: 18441233]

70. Bates PA, Kelley LA, MacCallum RM, Sternberg MJ. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins. 2001; (Suppl 5):39–46. [PubMed: 11835480]

71. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res. 2003; 31(13):3381–3385. [PubMed: 12824332]

72. Molecular Operating Environment 2004.03. Chemical Computing Group Inc; 1010 Sherbrooke Street West, #910, Montreal, Quebec, Canada H3A 2R7: 2004.

73. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. Proteins. 2007; 69(Suppl 8):108–117. [PubMed: 17894355]

74. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008; 9:40. [PubMed: 18215316]

75. Wallner B, Larsson P, Elofsson A. Pcons.net: protein structure prediction meta server. Nucleic Acids Res. 2007; 35:W369–374. Web Server issue. [PubMed: 17584798]

76. Chivian D, Kim DE, Malmström L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D. Automated prediction of CASP-5 structures using the Robetta server. Proteins. 2003; 53(Suppl 6):524–533. [PubMed: 14579342]

77. Pandit SB, Zhang Y, Skolnick J. TASSER-Lite: An automated tool for protein comparative modeling. Biophys J. 2006; 91(11):4180–4190. [PubMed: 16963505]

78. Fernandez-Fuentes N, Madrid-Aliste CJ, Rai BK, Fajardo JE, Fiser A. M4T: a comparative protein structure modeling server. Nucleic Acids Res. 2007; 35:W363–368. Web Server issue. [PubMed: 17517764]

79. Montgomerie S, Cruz JA, Shrivastava S, Arndt D, Berjanskii M, Wishart DS. PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. Nucleic Acids Res. 2008; 36:W202–209. Web Server issue. [PubMed: 18483082]

80. Fernandez-Fuentes N, Querol E, Aviles FX, Sternberg MJE, Oliva B. Prediction of the conformation and geometry of loops in globular proteins: Testing ArchDB, a structural classification of loops. Proteins. 2005; 60(4):746–757. [PubMed: 16021623]

81. Sellers BD, Zhu K, Zhao S, Friesner RA, Jacobson MP. Toward better refinement of comparative models: Predicting loops in inexact environments. Proteins. 2008; 72(3):959–971. [PubMed: 18300241]

82. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Sci. 2000; 9(9):1753–1773. [PubMed: 11045621]

83. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B. Loop modeling: Sampling, filtering, and scoring. Proteins. 2008; 70(3):834–843. [PubMed: 17729286]
84. Peng HP, Yang AS. Modeling protein loops with knowledge-based prediction of sequence-structure alignment. Bioinformatics. 2007; 23(21):2836–2842. [PubMed: 17827204]
85. Espadaler J, Fernandez-Fuentes N, Hermoso A, Querol E, Aviles FX, Sternberg MJ, Oliva B. ArchDB: automated protein loop classification as a tool for structural genomics. Nucleic Acids Res. 2004; 32:D185–188. Database issue. [PubMed: 14681390]
86. Fernandez-Fuentes N, Zhai J, Fiser A. ArchPRED: a template based loop structure prediction server. Nucleic Acids Res. 2006; 34:W173–176. Web Server issue. [PubMed: 16844985]
87. Lee DS, Seok C, Lee J. Protein loop modeling using fragment assembly. J Korean Phys Soc. 2008; 52(4):1137–1142.
88. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins. 2004; 55(2):351–367. [PubMed: 15048827]
89. Cui M, Mezei M, Osman R. Prediction of protein loop structures using a local move Monte Carlo approach and a grid-based force field. Protein Eng Des Sel. 2008; 21(12):729–735. [PubMed: 18957407]
90. Felts AK, Gallicchio E, Chekmarev D, Paris KA, Friesner RA, Levy RM. Prediction of protein loop conformations using the AGBNP implicit solvent model and torsion angle sampling. J Chem Theory Comput. 2008; 4(5):855–868. [PubMed: 18787648]
91. Xiang ZX, Soto CS, Honig B. Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci USA. 2002; 99(11):7432–7437. [PubMed: 12032300]
92. Olson MA, Feig M, Brooks CL. Prediction of protein loop conformations using multiscale Modeling methods with physical energy scoring functions. J Comput Chem. 2008; 29(5):820–831. [PubMed: 17876760]
93. Zhu K, Pincus DL, Zhao SW, Friesner RA. Long loop prediction using the protein local optimization program. Proteins. 2006; 65(2):438–452. [PubMed: 16927380]
94. Spassov VZ, Flook PK, Yan L. LOOPER: a molecular mechanics-based algorithm for protein loop prediction. Protein Eng Des Sel. 2008; 21(2):91–100. [PubMed: 18194981]
95. Zhu K, Shirts MR, Friesner RA. Improved methods for side chain and loop predictions via the protein local optimization program: Variable dielectric model for implicitly improving the treatment of polarization effects. J Chem Theory Comput. 2007; 3(6):2108–2119. [PubMed: 26636204]
96. Brooijmans N, Kuntz ID. Molecular recognition and docking algorithms. Ann Rev Biophys Biomol Struct. 2003; 32:335–373. [PubMed: 12574069]
97. Al-Lazikani B, Jung J, Xiang Z, Honig B. Protein Structure Prediction. Curr Opin Chem Biol. 2001; 5(1):51–56. [PubMed: 11166648]
98. Yanover C, Schueler-Furman O, Weiss Y. Minimizing and learning energy functions for side-chain prediction. J Comput Biol. 2008; 15(7):899–911. [PubMed: 18707538]
99. Lu MY, Dousis AD, Ma JP. OPUS-Rota: A fast and accurate method for side-chain modeling. Protein Sci. 2008; 17(9):1576–1585. [PubMed: 18556476]
100. Zhang W, Duan Y. Grow to Fit Molecular Dynamics (G2FMD): an ab initio method for protein side-chain assignment and refinement. Protein Eng Des Sel. 2006; 19(2):55–65. [PubMed: 16401632]
101. Yan A, Kloczkowski A, Hofmann H, Jernigan RL. Prediction of side chain Orientations in proteins by statistical machine learning methods. J Biomol Struct Dyn. 2007; 25(3):275–287. [PubMed: 17937489]
102. Canutescu AA, Shelenkov AA, Dunbrack RL. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci. 2003; 12(9):2001–2014. [PubMed: 12930999]
103. Hartmann C, Antes I, Lengauer T. IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. Protein Sci. 2007; 16(7):1294–1307. [PubMed: 17567749]

104. Spassov VZ, Yan L, Flook PK. The dominant role of side-chain backbone interactions in structural realization of amino acid code. ChiRotor: A side-chain prediction algorithm based on side-chain backbone interactions. Protein Sci. 2007; 16(3):494–506. [PubMed: 17242380]

105. Xie W, Sahinidis NV. Residue-rotamer-reduction algorithm for the protein side-chain conformation problem. Bioinformatics. 2006; 22(2):188–194. [PubMed: 16278239]

106. Dantas G, Corrent C, Reichow SL, Havranek JJ, Eletr ZM, Isern NG, Kuhlman B, Varani G, Merritt EA, Baker D. High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. J Mol Bio. 2007; 366(4):1209–1221. [PubMed: 17196978]

107. Lopes A, Alexandrov A, Bathelt C, Archontis G, Simonson T. Computational sidechain placement and protein mutagenesis with implicit solvent models. Proteins. 2007; 67(4):853–867. [PubMed: 17348031]

108. McGuffin LJ. The ModFOLD server for the quality assessment of protein structural models. Bioinformatics. 2008; 24(4):586–587. [PubMed: 18184684]

109. Gao X, Bu D, Li SC, Xu J, Li M. FragQA: predicting local fragment quality of a sequence-structure alignment. Genome Inform. 2007; 19:27–39. [PubMed: 18546502]

110. Wallner B, Elofsson A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci. 2006; 15(4):900–913. [PubMed: 16522791]

111. Mereghetti P, Ganadu ML, Papaleo E, Fantucci P, De Gioia L. Validation of protein models by a neural network approach. BMC Bioinformatics. 2008; 9:66. [PubMed: 18230168]

112. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 2007; 35:W407–410. Web Server issue. [PubMed: 17517781]

113. Benkert P, Tosatto SC, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins. 2008; 71(1):261–277. [PubMed: 17932912]

114. Eramian D, Shen MY, Devos D, Melo F, Sali A, Marti-Renom MA. A composite score for predicting errors in protein structure models. Protein Sci. 2006; 15(7):1653–1666. [PubMed: 16751606]

115. Qiu J, Sheffler W, Baker D, Noble WS. Ranking predicted protein structures with support vector regression. Proteins. 2008; 71(3):1175–1182. [PubMed: 18004754]

116. Randall A, Baldi P. SELECTpro: effective protein model selection using a structure-based energy function resistant to BLUNDERs. BMC Struct Biol. 2008; 8(1):52. [PubMed: 19055744]

117. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins. 2003; 53(Suppl 6):491–496. [PubMed: 14579338]

118. Paluszewski M, Karplus K. Model quality assessment using distance constraints from alignments. Proteins. 2009; 75(3):540–549. [PubMed: 19003987]

119. Costantini S, Facchiano AM, Colonna G. Evaluation of the structural quality of modeled proteins by using globularity criteria. BMC Struct Biol. 2007; 7:9. [PubMed: 17346357]

120. McGuffin LJ. Benchmarking consensus model quality assessment for protein fold recognition. BMC Bioinformatics. 2007; 8:345. [PubMed: 17877795]

121. Pettitt CS, McGuffin LJ, Jones DT. Improving sequence-based fold recognition by using 3D model quality assessment. Bioinformatics. 2005; 21(17):3509–3515. [PubMed: 15955780]

122. Wallner B, Fang H, Elofsson A. Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. Proteins. 2003; 53(Suppl 6):534–541. [PubMed: 14579343]

123. Colubri A, Jha AK, Shen MY, Sali A, Berry RS, Sosnick TR, Freed KF. Minimalist representations and the importance of nearest neighbor effects in protein folding simulations. J Mol Bio. 2006; 363(4):835–857. [PubMed: 16982067]

124. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006; 15(11):2507–2524. [PubMed: 17075131]

125. Tosatto SC. The victor/FRST function for model quality estimation. J Comput Biol. 2005; 12(10): 1316–1327. [PubMed: 16379537]

126. Wang Z, Tegge AN, Cheng J. Evaluating the absolute quality of a single protein model using structural features and support vector machines. Proteins. 2008; 75(3):638–647.

127. Zhou H, Skolnick J. Protein model quality assessment prediction by combining fragment comparisons and a consensus C(α) contact potential. Proteins. 2008; 71(3):1211–1218. [PubMed: 18004783]

128. Zhou HY, Zhou YQ. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 2002; 11(11): 2714–2726. [PubMed: 12381853]

129. Fasnacht M, Zhu J, Honig B. Local quality assessment in homology models using statistical potentials and support vector machines. Protein Sci. 2007; 16(8):1557–1568. [PubMed: 17600147]

130. Chen H, Kihara D. Estimating quality of template-based protein models by alignment stability. Proteins. 2008; 71(3):1255–1274. [PubMed: 18041762]

131. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Fams-ace: a combined method to select the best model after remodeling all server models. Proteins. 2007; 69(Suppl 8):98–107. [PubMed: 17894329]

132. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM. MetaMQAP: a meta-server for the quality assessment of protein models. BMC Bioinformatics. 2008; 9:403. [PubMed: 18823532]

133. Archie J, Karplus K. Applying undertaker cost functions to model quality assessment. Proteins. 2009; 75(3):550–555. [PubMed: 19004017]

34. Cavasotto CN, Phatak SS. Homology modeling in drug discovery: current trends and applications. Drug Discov Today. 2009; 14(13–14):676–683. [PubMed: 19422931]

135. Li N, Wang F, Niu S, Cao J, Wu K, Li Y, Yin N, Zhang X, Zhu W, Yin Y. Discovery of novel inhibitors of Streptococcus pneumoniae based on the virtual screening with the homology-modeled structure of histidine kinase (VicK). BMC Microbiol. 2009; 9:129. [PubMed: 19558698]

136. Innocenti A, Hall RA, Schlicker C, Scozzafava A, Steegborn C, Mühlschlegel FA, Supuran CT. Carbonic anhydrase inhibitors. Inhibition and homology modeling studies of the fungal β-carbonic anhydrase from Candida albicans with sulfonamides. Bioorg Med Chem. 2009; 17(13): 4503–4509. [PubMed: 19450983]

137. Meng X-Y, Zheng Q-C, Zhang H-X. A comparative analysis of binding sites between mouse CYP2C38 and CYP2C39 based on homology modeling, molecular dynamics simulation and docking studies. Biochim Biophys Acta. 2009; 1794(7):1066–1072. [PubMed: 19358898]

138. Sharon A, Chu CK. Understanding the molecular basis of HBV drug resistance by molecular modeling. Antiviral Res. 2008; 80(3):339–353. [PubMed: 18765256]

139. Gagnidze K, Sachchidanand, Rozenfeld R, Mezei M, Zhou M-M, Devi LA. Homology Modeling and Site-Directed Mutagenesis To Identify Selective Inhibitors of Endothelin-Converting Enzyme-2. J Med Chem. 2008; 51(12):3378–3387. [PubMed: 18507370]

140. Kiss R, Kiss B, Könczöl A, Szalai F, Jelinek I, László V, Noszál B, Falus A, Keseru GM. Discovery of Novel Human Histamine H4 Receptor Ligands by Large-Scale Structure-Based Virtual Screening. J Med Chem. 2008; 51(11):3145–3153. [PubMed: 18459760]

141. Patny A, Desai PV, Avery MA. Ligand-supported homology modeling of the human angiotensin II type 1 (AT(1)) receptor: insights into the molecular determinants of telmisartan binding. Proteins. 2006; 65(4):824–842. [PubMed: 17034041]

142. Evers A, Klabunde T. Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. J Med Chem. 2005; 48(4):1088–1097. [PubMed: 15715476]

143. Ekonomiuk D, Su XC, Ozawa K, Bodenreider C, Lim SP, Otting G, Huang D, Caflisch A. Flaviviral protease inhibitors identified by fragment-based library docking into a structure generated by molecular dynamics. J Med Chem. 2009; 52(15):4860–4868. [PubMed: 19572550]

144. Hritz J, de Ruiter A, Oostenbrink C. Impact of plasticity and flexibility on docking results for cytochrome P450 2D6: a combined approach of molecular dynamics and ligand docking. J Med Chem. 2008; 51(23):7469–7477. [PubMed: 18998665]

145. Rueda M, Bottegoni G, Abagyan R. Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. J Chem Inf Model. 2009; 49(3):716–725. [PubMed: 19434904]

146. Cavasotto CN, Orry AJ, Murgolo NJ, Czarniecki MF, Kocsi SA, Hawes BE, O'Neill KA, Hine H, Burton MS, Voigt JH, Abagyan RA, Bayne ML, Monsma FJ Jr. Discovery of novel chemotypes to a G-protein-coupled receptor through ligand-steered homology modeling and structure-based virtual screening. J Med Chem. 2008; 51(3):581–588. [PubMed: 18198821]

147. Nowak M, Kolaczkowski M, Pawlowski M, Bojarski AJ. Homology modeling of the serotonin 5-HT1A receptor using automated docking of bioactive compounds with defined geometry. J Med Chem. 2006; 49(1):205–214. [PubMed: 16392805]

148. Wang Q, Canutescu AA, Dunbrack RL Jr. SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. Nat Protoc. 2008; 3(12):1832–1847. [PubMed: 18989261]

149. Santana R, Larranaga P, Lozano JA. Side chain placement using estimation of distribution algorithms. Artif Intell Med. 2007; 39(1):49–63. [PubMed: 16854574]

150. Jain T, Cerutti DS, McCammon JA. Configurational-bias sampling technique for predicting side-chain conformations in proteins. Protein Sci. 2006; 15(9):2029–2039. [PubMed: 16943441]

151. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 2004; 32(suppl_2):W526–531. [PubMed: 15215442]

Table 1

Summary of recent advances of protein structure prediction methodology.

Method	Web Address	Reference	Availability
Secondary structure prediction			
IPSSP	http://exon.gatech.edu/ipssp/webIPSSP.cgi	[57]	Server
MUPRED	http://digbio.missouri.edu/mupred	[53]	Server/Download
CDM	http://gor.bb.iastate.edu/cdm	[51]	Server
Jpred	http://www.compbio.dundee.ac.uk/jpred	[52]	Server
YASSPP	http://glaros.dtc.umn.edu/yasspp	[54]	Server
Proteus	http://wks16338.biology.ualberta.ca/proteus	[49, 50]	Server
DISTILL-Porter(_H)	http://distill.ucd.ie/distill	[49]	Server
P.S.HMM	http://nash.ucsd.edu/P_single.html	[61]	Server
DBNN	http://ctb.pku.edu.cn/main/SheGroup/Software/DBNN	[56]	Download
E-SSpred	http://bioinfo.hust.edu.cn/bio/tools/E-SSpred/index.html	[62]	Server
Fold recognition/remote homology detection			
Phyre	http://www.sbg.bio.ic.ac.uk/~phyre	[21]	Server
PDBalert	http://toolkit.tuebingen.mpg.de/pdbalert	[22]	Server
SP4	http://sparks.informatics.iupui.edu/SP4	[24]	Server
SP5	http://sparks.informatics.iupui.edu/SP5	[25]	Server
FoldPro	http://mine5.ics.uci.edu:1026/foldpro.html	[26]	Server
COMPASS	http://prodata.swmed.edu/compass/compass.php	[27]	Server
HMM-Kalign	http://www-spider.cea.fr/Groups/hk3039/view.html	[41]	Download
Loop Prediction			
ArchPRED	http://www.fiserlab.org/servers/archpred	[86]	Server
MMC	http://atlas.physbio.mssm.edu/~mezei/mmc	[89]	Download
PrISM	http://cmb.genomics.sinica.edu.tw	[84]	Server
Side-Chain Prediction			
OPUS-ROTA	http://sigler.bioch.bcm.tmc.edu/MaLab	[99]	Download
SCWRL and MolIDE	http://dunbrack.fccc.edu/Software.php	[148]	Download
R3	http://eudoxus.scs.uiuc.edu/r3.html	[105]	Server
SPRINT	http://www.protonet.cs.huji.ac.il/sprint	[149]	Download
IRECS	http://irecs.bioinf.mpiinf.mpg.de	[103]	Download
SPRUCE[§]	http://mccammon.ucsd.edu	[150]	Download[§]
Structure prediction			
I-TASSER	http://zhang.bioinformatics.ku.edu/I-TASSER	[73]	Server
Pcons	http://pcons.net	[75]	Server/download
ROBETTA	http://robetta.bakerlab.org/submit.jsp	[151]	Server/download
M4T	http://www.fiserlab.org/servers/m4t	[78]	Server
TASSER-Lite	http://cssb.biology.gatech.edu/skolnick/webservice/tasserlite/index.html	[77]	Server

Method	Web Address	Reference	Availability
Assessment of model quality			
Undertaker (SAM_T08)#	http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html	[118, 133]	Server#
QMEAN	http://swissmodel.expasy.org/qmean/cgi/index.cgi	[113]	Server
SPAD	http://dragon.bio.purdue.edu/subalignment	[130]	Software
–	http://bioinformatica.isa.cnr.it/GLOBULARITY	[119]	Analysis
ModFOLD	http://www.reading.ac.uk/bioinf/ModFOLD	[108]	Server
Selectpro	http://www.igb.uci.edu/~baldig/selectpro.html	[116]	Server/Download
ProQres/ProQprof	http://www.sbc.su.se/~bjorn/Pro Q	[110]	Server
ModelEvaluator	http://babbage.cs.missouri.edu/~chengji/cheng_software.html	[126]	Software
ProSA-web	https://prosa.services.came.sbg.ac.at/prosa.php	[112]	Server
MetaMQAP	https://genesilico.pl/toolkit	[132]	Server
Fams-ace (FAMS)#	http://www.pharm.kitasato-u.ac.jp/fams	[131]	Server

#Part of automated homology modeling server listed in parentheses;

§Made available by original authors upon request