Arch. Biol. Sci., Belgrade, 62 (4), 857-871, 2010 DOI:10.2298/ABS1004857M

BASIC PROTEIN STRUCTURE PREDICTION FOR THE BIOLOGIST: A REVIEW

M. MIHĂŞAN

"Alexandru Ioan Cuza" University, Faculty of Biology, Department of Molecular and Experimental Biology,
6600 Iaşi, Romania

Abstract - As the field of protein structure prediction continues to expand at an exponential rate, the bench-biologist might feel overwhelmed by the sheer range of available applications. This review presents the three main approaches in computational structure prediction from a non-bioinformatician's point of view and makes a selection of tools and servers freely available. These tools are evaluated from several aspects, such as number of citations, ease of usage and quality of the results. Finally, the applications of models generated by computational structure prediction are discussed.

Key words: Protein structure prediction, protein mode application

UDC 57.081.088.6

INTRODUCTION

Knowledge of the three-dimensional structure of a protein can provide invaluable hints about its functional and evolutionary features and, in addition, the structural information is useful in drug design efforts. Genome-scale sequencing projects have already produced more than 108 million individual sequences (Benson et al., 2010), but due to the inherently time-consuming and complicated nature of structure determination techniques, only around 53,000 of these have their 3D structures solved experimentally (Dutta et al., 2009). In spite of great progress in structural genomics, it is still unreasonable to believe that the structure of more than a tiny fraction of all the billions of proteins will be studied by experimental methods in the foreseeable future (Wallner and Elofsson, 2005). This places computer-based protein structure prediction in an unprecedentedly important position as the only reasonable means to bridge the gap between the number of known sequences and that of 3D models. The performance of *in silico* methods of protein structure prediction has recently improved significantly and dozens of servers and stand-alone programs are currently available (Fischer, 2006). This is evident from a PubMed query using the terms "protein structure prediction AND server". The query returns 388 articles, of which more than half were published in the past three years alone, the rest being published between 1993 and 2006. Because of this proliferation, it is difficult for a biologist to know which server or program to use, as it is hard to answer frequent questions such as: how do I know if I can trust the result; what does output mean; should I use more than one server; how much time will it take to get results? This review will address these questions with particular emphasis on the evaluation of the currently available free programs and web-servers from a biologist's point of view.

Protein structure prediction methods

According to Anfinsen's (1973) thermodynamic hypothesis, proteins are not assembled into their native structures by a biological process. Protein folding is a purely physical process that depends only on the specific amino acid sequence of the protein and the surrounding solvent (Anfinsen, 1973). This would suggest that one should be able to predict, at least theoretically, the three-dimensional (3D) conformation of a protein from its sequence alone. Since then, many efforts have been devoted to this fascinating and challenging problem, attempting to tackle this problem from different angles including biophysics, chemistry, and

biological evolution. Solving the problem of predicting a protein's 3D structure from its amino acid sequence has been called the "holy grail of molecular biology" and is considered as equivalent to deciphering "the second half of the genetic code" (Kolata, 1986).

The study of the principles that dictate the 3D structure of natural proteins can be approached either through the laws of physics or the theory of evolution. Each of these approaches provides the foundation for a class of protein structure prediction methods (Fiser, 2004). Accordingly, theoretical structure prediction can be divided into two extreme camps: homology modeling and *ab initio* methods (Xiang, 2006). The boundaries between these two extreme classes of prediction techniques have started to become blurred as scientists have started to integrate the strengths of different methods to make their prediction methods more effective and more generally applicable. Also, a third class of protein structure prediction methods has appeared: protein threading.

Homology modeling makes structure predictions based primarily on its sequence similarity to one or more proteins of known structures. *Ab initio* methods predict the three-dimensional structure of a given protein sequence without using any structural information of previously solved protein structures; instead, methods belonging to this group are entirely based on the first principles of physics (Pillardy et al., 2001). Protein threading, sometimes referred as fold recognition (FR) is an approach between the two extremes which uses both sequence similarity information when it exists, and structural fitness information between the query protein and the template structure (Jun-tao, Kyle and Ying, 2008). Below is a brief discussion of each of these methods, which emphasizes their advantages and disadvantages from an user's point of view.

Homology modeling, also referred to as comparative modeling (CM), is a class of methods based on the fact that proteins with similar sequences adopt similar structures, as most protein pairs with more than 30 out of 100 identical residues were found to be structurally similar (Rost,1999). Homology modeling is facilitated by the fact that the 3D structure of proteins from the same family is more conserved than their amino acid sequences (Lesk and Chothia, 1980). When the structure of one protein in a family has been determined by experimentation, other members of the same family can be modeled based on their alignment to the known structure. This high robustness of structures with respect to residue exchanges explains partly the robustness of organisms with respect to gene-replication errors, and it allows for the variety in evolution.

Comparative modeling consists of five main stages: (a) identification of evolutionary related sequences of known structure; (b) aligning of the target sequence to the template structures; (c) modeling of structurally conserved regions using known templates; (d) modeling side chains and loops which are different than the templates; (e) refining and evaluating the quality of the model through conformational sampling (Floudas, 2007). The accuracy of predictions by homology modeling depends on the degree of sequence similarity between the target sequence and the template structures. When the sequence identity is above 40%, the alignment is straightforward, there are not many gaps, and 90% of main-chain atoms could be modeled with an RMSD (root-mean-square distance) error of about 1 Å (Xiang, 2006). In this range of sequence identity, predictions are of very good to high quality, and have been shown to be as accurate as low-resolution X-ray predictions (Kopp and Schwede, 2004).

When the sequence identity is about 30-40%, obtaining correct alignment becomes difficult where insertions and deletions are frequent. For sequence similarity in this range, 80% of main-chain backbone atoms can be predicted to RMSD 3.5 Å, while the rest of the residues are modeled with larger errors (Xiang, 2006).

When the sequence identity is below 30%, the main problem becomes the identification of the homolog structures, and alignment becomes much

more difficult. Even if positive hits are found, their significance is questionable, thereby giving rise to the name of the 20 -30 % zone – the twilight zone of protein sequence alignments (Rost, 1999).

From a user point of view, the main difficulty in homology modeling is finding the target sequence to be used as a template. Approximately 57% of all known sequences have at least one domain that is related to at least one protein of known structure (Pieper et al., 2002). The probability of finding a related known structure for a randomly selected sequence from a genome ranges from 30% to 65% (Xiang, 2006). The percentage is steadily increasing because projects like Protein Structure Initiative promise to fulfill within the next decade (Zhang, 2009b) the task of experimentally determining the 16 000 optimally selected new structures needed so that homology modeling can cover 90% of protein domains (Vitkup et al., 2001).

Protein threading

Also known as fold recognition (FR), protein threading is a class of methods that aims at fitting a target sequence to a known structure in a library of folds. Generally, similar sequence implies similar structure but the converse is not true: similar structures are often found for proteins for which no sequence similarity to any known structure can be detected (Floudas et al., 2006). This means that the actual number of different folded protein structures is significantly smaller than the number of different sequences generated by the large scale genome projects (Floudas, 2007). An optimistic view is that the number of existing folds is a few orders of magnitudes smaller than the number of different sequences, possibly ranging from a few hundred to a few thousand.

The basic idea of protein threading is to literally "thread" the amino acids of a query protein, following their sequential order and allowing for insertions and gaps, into the structural positions of a template structure in an optimal way measured by a scoring function. This procedure is repeated for each template structure in a database of protein folds. The quality of a sequence-structure alignment is typically assessed using statistical-based energy and the "best" sequence-structure alignment provides a prediction of the backbone atoms of the query protein.

The main drawback of this class of methods is the fact that it is very demanding on the computing power and also, that there is still a need for target identification. Currently, the Protein Data Bank contains enough structures to cover small single-domain protein structures up to a length of about 100 residues, so the method has the best chances of success with proteins within this limit (Kihara and Skolnick, 2003; Zhang and Skolnick, 2005).

Ab initio methods

Also known as *de novo* methods, "first principle" methods or "free modeling" (Zhang, 2008b), these methods assume that the native structure corresponds to the global free energy minimum accessible during the lifespan of the protein, and attempt to find this minimum by an exploration of many conceivable protein conformations (Fiser, 2004). The term *ab initio* methods referred initially to methods for structure prediction that do not use experimentally known structures (Floudas et al., 2006). Lately, this term has become vaguer since the introduction of novel fragment based methods. These methods primarily utilize the fact that, although we are far from observing all folds used in biology (Coulson and Moult, 2002), we probably have seen nearly all substructures (Du Andrec and Levy, 2003). Structure fragments are chosen on the basis of the compatibility of the substructure with the local target sequence and assembled into one new structure. The field of *ab initio* prediction methods is thereby divided into two main classes: *ab initio* methods with database information and *ab initio* methods without database information (Floudas et al., 2006).

Even though the methods from this last class are computationally very demanding and still lack accuracy (Fiser, 2004), they are continuously used and developed for several reasons. Firstly, in some cases, even a remotely related structural homolog may not be available. In these cases, *ab initio* methods are the only alternative. Secondly, new struc-

tures continue to be discovered which could not have been identified by methods which rely on comparison to known structures. Thirdly, knowledge-based methods have been criticized for predicting protein structures without having to obtain a fundamental understanding of the mechanisms and driving forces of structure formation. First principle structure prediction methods, in contrast, base their predictions on physical models for these mechanisms. As such, they can therefore help to deepen the understanding of the mechanisms of protein folding (Floudas et al., 2006).

From a user point of view the main bottlenecks of *ab initio* methods are the resolution of generated models and the computing power required to generate these models. The low resolution of *ab initio* generated models resides in our limited understanding of the protein folding problem and despite significant progress in this direction (Bonneau and Baker, 2001), it remains applicable to a limited number of sequences of less than approximately 100 residues (Fiser, 2004).

Programs, servers and meta-servers

Because computer-based protein structure prediction methods have so much to offer, the scientific community has invested a tremendous effort into solving the different problems and bottlenecks that each method has. Dozens of ingenious solutions have emerged and a dizzy array of methods is implemented in various tools. The sheer range of available tools can be overwhelming for the bench-work biologist who might find it hard to chose the right one for the job.

These tools can be classified from a computation point of view as a stand-alone program, a server or a meta-server. In the past, *in-silico* protein structure prediction was invariably performed using stand-alone programs such as: What If, SegMod/ENCAD, nest or builder (Table 1). This required both skills in different programming languages as well as access to high computing power. This explains why protein structure prediction was previously performed only by a handful of specialized researchers. Today, structural information is required by a increasingly large and diverse group of scientists and most of them are not prepared to spend months learning the complex user interfaces of various operating systems or complicated scripting languages.

Web-servers free the biologist from the burden of implementing and/or maintaining complicated and resource demanding software (Fischer, 2006). This is done by specialized bioinformaticians. All the computations are done elsewhere; the average user only submits his amino acid sequence via a web browser and then waits for the results, most of the time by e-mail. This approach has been a real breakthrough in terms of user-friendliness and has gained huge success lately. Servers such as Swiss-Modell have been cited no less than 253 times in various papers published in 2009 (Fig. 1). The use of such autonomous servers has a huge drawback: each server uses only one method of prediction with its corresponding flows. Two different servers, using different prediction methods, will give different results for the same query. So human predictors have still to improve the model manually, they have to determine which of the obtained model is correct, whether there is a lower ranking model that corresponds to a correct prediction, or whether the results of the method indicate that no prediction at all can be obtained.

One step forward in the full automation of the protein prediction process is the emergence of meta-servers. We distinguish meta-servers from autonomous servers by the type of input required: a meta-server cannot run independently, explicitly requiring as input the predictions of at least one other participating server (Bujnicki and Fischer, 2004). A meta-server doesn't make the prediction based on only one method, but combines the results from several other servers, each using its own methods of prediction. As in the case of autonomous servers, the user is required to simply input the amino acid sequence in a web browser. The meta-server will then run the sequence through several other servers, obtain and rank the results and then send the final results back to the user (most of the time by e-mail).

Ease of use is one important aspect from a user point of view reflected directly in the number of

Table 1. Selected programs and servers for proteins structure prediction

Name	Comments	URL
	Stand-alone	
	Homology modeling	
CABS (Kolinski, 2004)	*de novo* folding of small proteins, comparative modeling. Also accessible through the Selvita Protein Modeling Platform OS: Linux	www.biocomp.chem.uw.edu.pl/services.php
Fold-X (Schymkowitz et al., 2005)	commercial program, on site registration possible for a 15 days trial	http://foldx.crg.es/
Modeller (Fiser and Sali, 2003b; Sali et al.,1995; Sanchez and Sali,1997)	command line interface, GUIs and web-servers also available OS: Windows, Mac, Linux	www.salilab.org/modeller/
What If	commercial program	http://swift.cmbi.ru.nl/whatif/
nest	part of the JACKAL software package, combines template-based methods with *ab initio*-like energy minimization principles OS:SGI 6.5, Intel Linux and Sun solaris.	http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:Jackal
Biskit (Gruenberg et al., 2007)	a python library for structural bioinformatics research OS:Linux, Windows	http://biskit.pasteur.fr/
	Threading/fold recognition	
SUPERFAMILY (Gough et al., 2001; Madera et al., 2004; Wilson et al.,2007; Wilson et al., 2009)	Hidden Markov modeling	http://supfam.org/SUPERFAMILY/
	Servers	
	Homology modeling	
3D-JIGSAW (Bates et al., 2001)	fully automated system which can be also run in interactive mode	www.bmm.icnet.uk/~3djigsaw/
EsyPred3D (Lambert et al., 2002)	automated homology modeling, The final three dimensional structure is built using the modeling package MODELLER	http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/
Geno3D (Combet et al., 2002)	comparative protein structure modeling by spatial restraints satisfaction, generates models containing up to 500 amino acids	http://geno3d-pbil.ibcp.fr
HHPred (Sodingt et al., 2005)	homology detection and structure prediction by HMM-HMM comparison	http://toolkit.tuebingen.mpg.de/hhpred#
HHPred (Sodingt et al., 2005)	homology detection and structure prediction by HMM-HMM comparison	http://toolkit.tuebingen.mpg.de/hhpred#
HHPred (Sodingt et al., 2005)	homology detection and structure prediction by HMM-HMM comparison	http://toolkit.tuebingen.mpg.de/hhpred#

Table 1. Continued.

Name	Comments	URL
HHPred (Sodingt et al., 2005)	homology detection and structure prediction by HMM-HMM comparison	http://toolkit.tuebingen.mpg.de/hhpred#
Swiss-Modell (Arnold et al., 2006a; Bordoli et al., 2009a; Kiefer et al., 2009; Peitsch, 1995)	fully automated protein structure homology-modeling server, accessible also from the program DeepView (Swiss Pdb-Viewer)	http://swissmodel.expasy.org/
	Threading/fold recognition	
3D-PSSM (Kelley et al., 2000)	Synce 2004 the development of this server has been frozen	http://www.sbg.bio.ic.ac.uk/~3dpssm/index2.html
Phyre (Kelley and Sternberg, 2009b)	The successor of 3D-PSSM	http://www.sbg.bio.ic.ac.uk/~phyre/
I-TASSER (Zhang, 2007; Zhang, 2008a; Zhang, 2009a)	3D models are built based on multiple-threading alignments by LOMETS and iterative TASSER simulations; was ranked as the No 1 server for protein structure prediction in recent CASP7 and CASP8 experiments (the "Zhang-Server")	http://zhanglab.ccmb.med.umich.edu/I-TASSER/
LOOPP (Meller and Elber, 2001; Teodorescu et al., 2004; Tobi and Elber, 2000)	fold recognition program based on the collection of numerous signals, merging them into a single score, and generating atomic coordinates based on an alignment into a homolog template structure	http://cbsuapps.tc.cornell.edu/loopp.aspx
Muster (Wu and Zhang, 2008)	it generate sequence-template alignments by combining sequence profile-profile alignment with multiple structural information	http://zhanglab.ccmb.med.umich.edu/MUSTER/
	Ab-initio	
ModLoop (Fiser et al., 2000; Fiser and Sali, 2003a)	automated modeling of loops in protein structures, relies on the loop modeling routine in MODELLER	http://modbase.compbio.ucsf.edu/modloop/
Phyre (Kelley and Sternberg, 2009b)	The successor of 3D-PSSM	http://www.sbg.bio.ic.ac.uk/~phyre/
	Metaserver	
Lomets (Wu and Zhang, 2007)	generates 3D models by collecting high-scoring target-to-template alignments from 8 locally-installed threading programs	http://zhanglab.ccmb.med.umich.edu/LOMETS/
GeneSilico (Kurowski and Bujnicki, 2003)	on-site registration required	www.genesilico.pl/meta2/
Meta-PP (Eyrich and Rost, 2003; Rost, 1996)	the job cab be submitted to up to 12 servers, among which 2 threading and two homology modeling servers	http://www.cs.bgu.ac.il/~dfischer/predictprotein/submit_meta.html
3D-JURY (Ginalski et al., 2003)	uses about 10 different servers for a prediction, among which 3D-PSSM and	http://meta.bioinfo.pl/submit_wizard.pl
Robetta (Chivian and Baker, 2006; Chivian et al., 2003; Kim et al., 2004)	homology modeling, ab initio structure prediction, and structure prediction using NMR constraints	http://robetta.bakerlab.org/

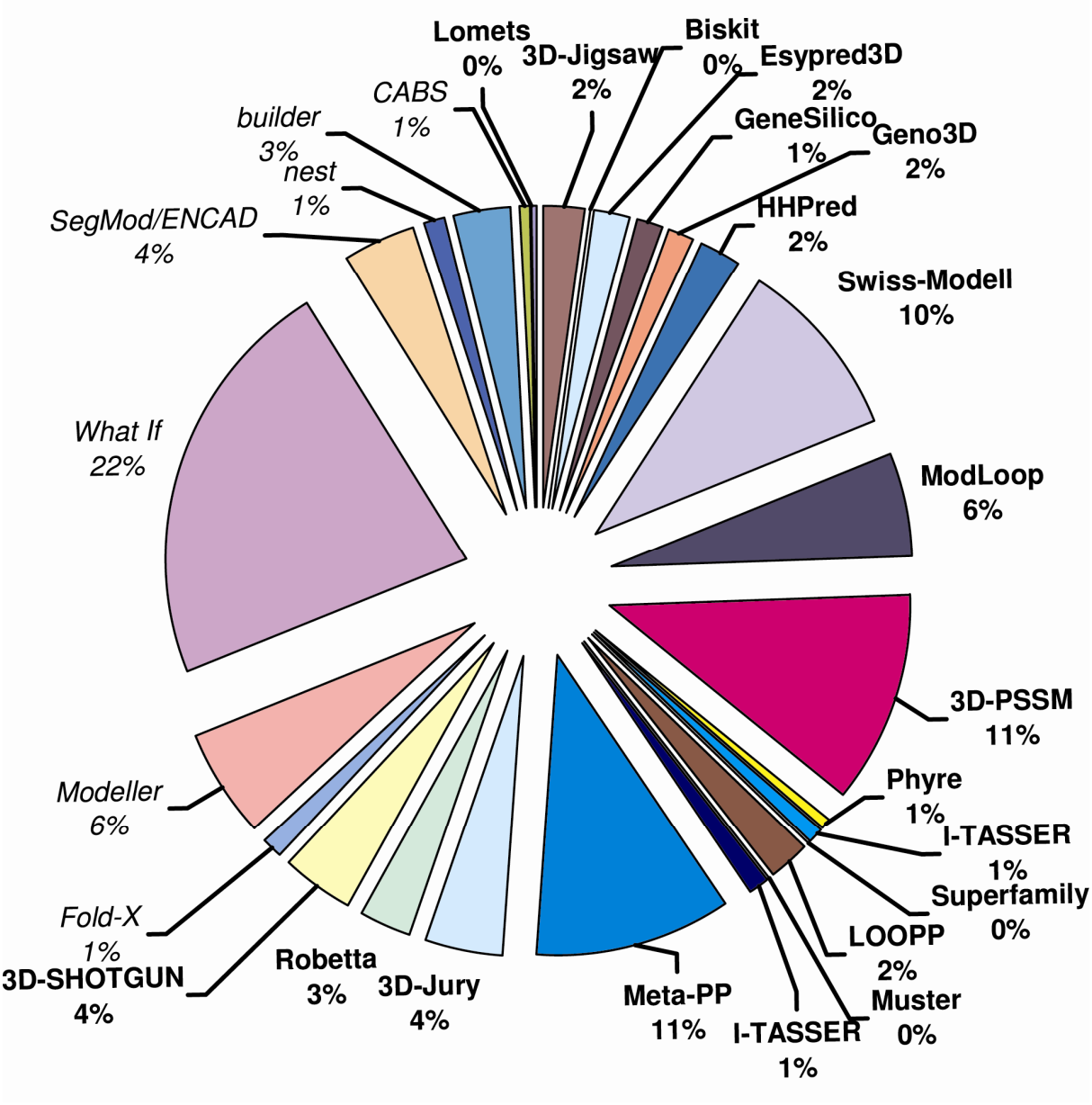

Figure 1. Protein structure prediction software – number of citations for some of the most common docking programs (italics) and servers (bold type), analyzed from the ISI Web of Science (2009) considering any of the original references as indicated in Table 1.

citations (Fig. 2). It is not by far the most important aspect. Model quality must be the key feature when deciding which service to use. As this cannot be easily inferred by the average biologist, projects like LiveBench (Rychlewski and Fischer, 2005), CASP (Critical Assessment of techniques for Structure Prediction) (Kryshtafovych et al., 2009) and EVA (Koh et al., 2003) focus on benchmarking all these programs, servers and meta-servers. In the world of predictors, the most famous is CASP – as some sort of Olympian in protein structure prediction. Human and computer predictors, or a mix of the activities in order to participate in a CASP round. Moreover, the Proteins journal publishes, since

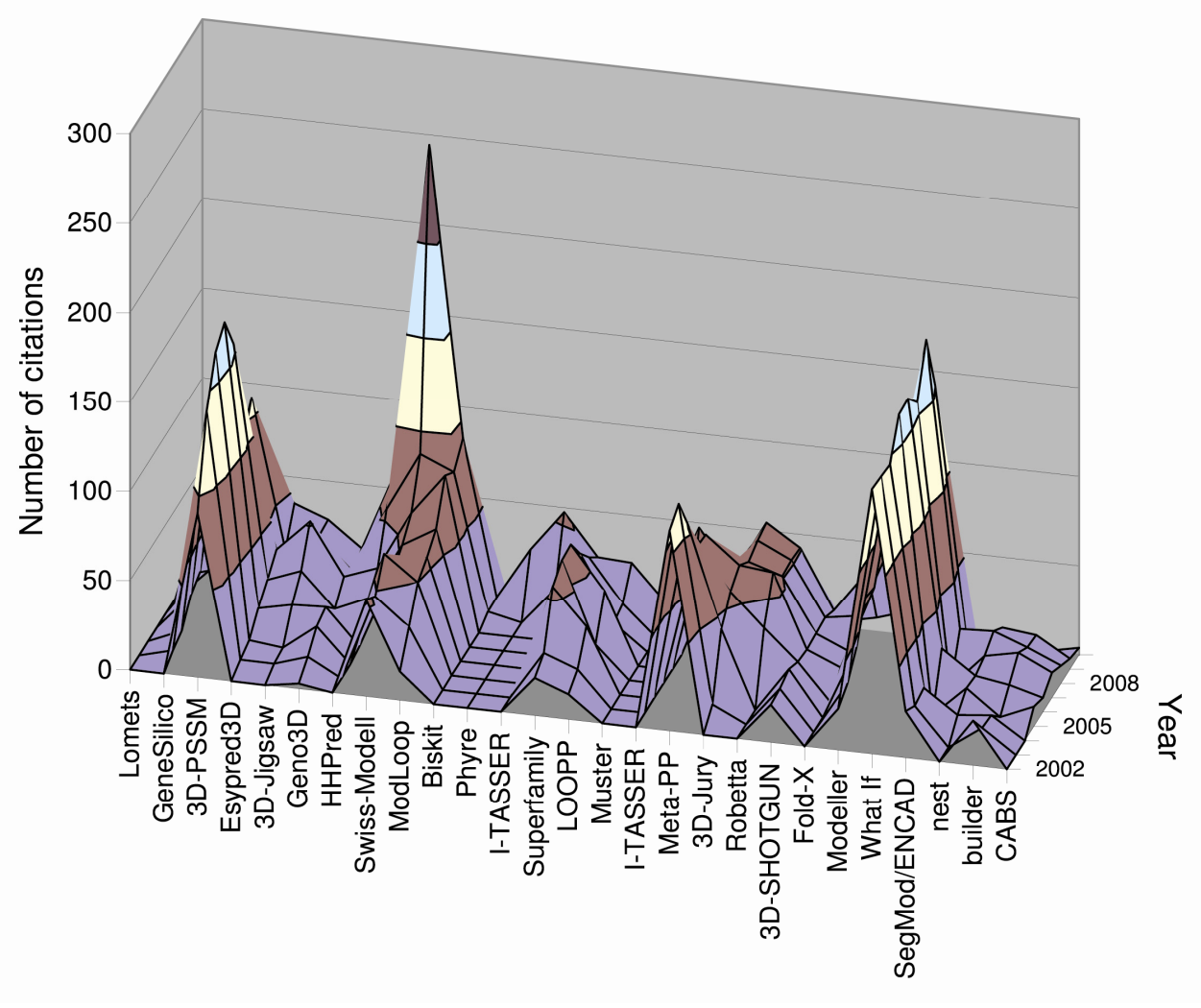

Figure 2. Protein structure prediction software – trends in the number of citations per year for some of the most common docking programs and servers, analyzed from the ISI Web of Science (2009) considering any of the original references as indicated in Table 1.

1995, a special issue on CASP every two years (Lattman, 1995). So CASP is definitely a place that must be periodically checked for the latest news in the field of protein prediction.

Model quality and evaluation

As a server or a program will almost always return an answer, using two or more of such tools means that one will get more than just one computer-generated model. How does one know which to choose, which he can trust most? As opposed to experimental structure evaluation, there are not too many reliable procedures to assess the quality of a computer-generated model (Petrey et al., 2003). Before tackling with any *in silico* protein prediction problem, a non-bioinformatician has to check the CASP website. Choosing a tool from most highly ranked in the latest CASP experiment will assure the best possible start in terms of reliability of the results.

Beside the CASP rank, another important factor in choosing the right tool is the protein to be modeled.

There is a basic rule to follow. If your protein has at least 40% similarity with a known structure, comparative modeling is the method to use. For lower similarities, protein threading is preferred. When the target sequence has no similarities with known structure, *ab initio* methods are the last resort.

Two types of evaluation of the computer-generated models can be carried out. Internal evaluation of self consistency checks whether or not a model satisfies the restraints used to calculate it. Generally, each of the tools used in the construction of a model, template selection, alignment, model building, and refinement has its own internal measures of quality (Petrey et al., 2003). Nevertheless, assessment of the stereochemistry of a model (e.g., bonds, bond angles, dihedral angles and non-bonded atom-atom distances) can be additionally checked with programs such as PROCHECK (Laskowski et al., 1993), WHAT-IF (Vriend, 1990) and WHAT-CHECK (Hooft et al., 1996).

External evaluation relies on information that was not used in the calculation of the model, like the calculation of the pseudo energy profile of a model performed by tools like PROSA (Sippl, 1995), Verify3D (Eisenberg et al., 1997) and QMEAN (Benkert et al., 2008).

Finally, a model should be consistent with any existing experimental observations, such as site-directed mutagenesis, cross-linking data and ligand binding (Fiser, 2004). The review of Kihara et al. (2009) is a very good starting point for further reading on the various errors frequently found in computer-generated models and different methods of detection.

Examples of widely used structure prediction tools

The Swiss-Model Workspace (Arnold et al., 2006b) can be freely accessed by the biological community on the Web at http://swissmodel.expasy.org/ workspace/. The Swiss-Model has been the first automated modeling server publicly available (Peitsch, 1995) and since then it has been cited no less than 896 times. It uses homology modeling as the prediction method and besides a very intuitive fully automated mode, it also has a project mode which allows the user to manually select the template and edit the alignment before modeling. Most importantly, the Swiss-Model includes several tools for structure assessment such as PROCHECK (Laskowski et al., 1993), WHAT-IF (Vriend, 1990) and QMEAN (Benkert et al., 2008), being in this case a very complete package.

All the tools are organized as a workspace, where the user logs-in using an e-mail address and a server-provided password. Once a modeling request is submitted, its status can be monitored on the workspace and when the job is finished and the user is notified by e-mail. The results are kept on the server for 7 days, the user being able to expand that period on choice. Bordoli et al. (2008) provide a step-by-step guide in protein modeling using Swiss-Model (Bordoli et al., 2009b).

3D-JIGSAW is an automated system to build three-dimensional models for proteins based on homologs of a known structure. This system is modular in design with each module centering on a particular algorithm required in the modeling process (Bates et al., 2001). The system can either be run locally or via a web server (http://www.bmm.icnet.uk/~3djigsaw/). In the web server version, the user inputs the sequence in one-letter code, fills-in the e-mail address and then has to choose a building mode: automatic or interactive.

In the automatic mode the server looks for homologous templates in several sequence databases and splits the query sequence into domains. If good templates are found, the best covered domain is then modeled. The process can take up to an hour, depending on the load of the system. The user will receive the alignment between query and template/s and a PDB formatted set of coordinates by e-mail.

In the interactive mode, the program looks for homologous templates in the sequence databases and splits the query sequence into domains. An e-

mail is sent back to the user with a link to a graphical display of this domain arrangement and useful information extracted from the PFAM database. From this link the user may chose the domains for modeling and may select the templates and the correct alignments before submitting a modeling job. Templates are ranked according to the coverage of the query, their sequence identity and their crystallographic resolution. Like in the automatic mode, the final results will be sent to the user by e-mail.

Up to now the 3D-JIGSAW system has reached version 2.0, version 3.0 being in the pre-release stage (http://www.bmm.icnet.uk/~populus/).

Modeller is a stand-alone command line program available for Unix/Linux, Windows and Mac systems, which implements comparative protein structure modeling by satisfaction of spatial restraints (Fiser et al., 2000; Sali & Blundell, 1993). The current version of this software is Moldeller 9v8, and is available free-of-charge to academic non-profit institutions and from Accelersys for commercial entities.

An example of comparative modeling using Modeller with some very detailed step-by-step instructions on using the command line interface is provided by Eswar et al., 2007 (Eswar et al., 2007). As the command line and Modeller control language could be found hard to learn for an average user, several graphical user interfaces such as Easy-Modeller and Mint have been developed by a third party and are freely available.

Robetta provides an on-line interface for the Rosetta protein modeling suite guided towards homology modeling, *ab initio* structure prediction and structure prediction using NMR constraints. Comparative models are built from Parent PDBs detected by UW-PDB-BLAST or 3DJury-A1 and aligned by the K*SYNC alignment method (Chivian and Baker, 2006; Kaufmann et al., 2010). Domains with no detectable PDB homolog are modeled with the Rosetta *de novo* protocol [(Bonneau et al. (2002); Simons et al., 1997)]. The procedure is fully automated and the server is only available for use by the academic community and other not-for-profit entities. In the last CASP experiment, the Rosetta server was ranked among the top-three servers. Some good guidelines on working with the Rosetta server are provided by Chivian et al. (2003) and Kim et al. (2004).

3d-pssm, although widely used, with a record 1039 citations at the time of writing of this manuscript, this protein threading server has been replaced by its successor, the new and improved Phyre server. Since its launch in 2002, the Phyre server has already been cited no less than 53 times, scoring very well in the CASP8 experiments. A detailed description of the protein structure prediction protocol with the Phyre server is provided by Kelley and Sternberg (2009a).

META-PP is a meta-server which provides a simple streamlined interface to a wide range of prediction servers in computational biology/bioinformatics. Users access the server via a simple web interface (http://cubic.bioc.columbia.edu/meta/). Input is a one-letter code protein sequence along with an optional short description of the protein and an email address. Users then manually select the sub-set of available servers they want to access. META-PP validates the input (email address and sequence format) and places the request into a processing queue. During the processing of a prediction request META-PP assembles the raw data required for submission, such as sequences and job options, connects to the remote server using the appropriate protocol and submits the request. Depending on the server, META-PP might wait and receive actual output in real-time or simply wait for submission confirmation and then disconnect. In the case of failure, caused, for example, by intermittent outages at the remote site or by simple connectivity problems, META-PP reinserts the failed request into its own processing queue and resubmits at a later time (for up to 24 h, after which failed prediction requests are simply purged from the processing queue). Depending on the characteristics of the prediction server, users will receive results either from META-PP or directly from the original prediction server (Eyrich & Rost, 2003).

Applications of structure predictions

A 3-D model does not have to be absolutely perfect to be helpful in biology, but the type of question that can be addressed with a particular model does depend on its accuracy. Depending on the prediction approach applied (Fiser, 2004) the accuracy of a model differs. Comparative modeling generates structures that have a root mean square deviation (RMSD) of 1–2 Å from the experimental structure, achieving the accuracy of medium-resolution NMR or low-resolution X-ray structures (Read and Chavali, 2007). Threading provides models with an RMSD of 2–6 Å, with errors mainly occurring in the loop regions (Jauch et al., 2007). For target proteins without solved template structures, *ab initio* methods are limited to small proteins (<120 residues) with an accuracy in the range of 4–8 A (Kopp et al., 2007). For low accuracy models (RMSD >*3* Å) RMSD is no longer a meaningful measure of modeling quality (Fiser, 2004) and TM-score is preferred. By definition, TM-score lies in a 0.1 interval. A TM value of 1 indicates a very accurate model (equivalent of RMSD 0 Å), a value >0.5 indicates a model with a roughly correct topology, and a value 0.17 indicates a random prediction regardless of the protein size (Zhang, 2009b).

High-resolution models obtained by homology modeling at more than 50% sequence identity can usually meet the highest structural requirements in the case of single-domain proteins and have been use in a wide range of applications, as docking, designing and improving ligands for a given binding site (Ring et al., 1993), designing mutants to test hypotheses about a protein's function (Vernal et al., 2002; Wu et al., 1999), identifying active and binding sites (Sheng et al., 1996), simulating protein-protein docking (Vakser, 1995), facilitating molecular replacement in X-ray structure determination (Howell et al., 1992), refining models based on NMR constraints (Modi et al., 1996) and rationalizing known experimental observations (Eswar et al., 2007).

For models of medium-resolution, with an RMSD between 2.5-5 Å, typically generated by comparative modeling from distantly homologous templates or by fold recognition, the structural predictions are useful for identification of the spatial locations of functionally important residues, such as active sites and the sites of disease-associated mutations. Arakaki et al. (2004) assessed the possibility of assigning the biological function of enzyme proteins by matching the structural patterns (or descriptors) of the active sites with structure decoys of various resolutions. Boyd et al. (2008) used structural models generated by the automated I-TASSER server to help interpret mutagenesis experiments with the Sec1/Munc18 (SM) proteins on the basis of the spatial clustering of the mutated residues.

Models with the lowest resolution from free modeling approaches or based on weak hits from threading, have a number of uses including protein domain boundary identification (Tress et al., 2007), topology recognition, or family/superfamily assignment. For example, the TASSER structural predictions placed the RDC1 receptor in the family of chemokine receptors because the predicted RDC1 structure is closest to the predicted structure of the CXCR4 chemokine receptor (Zhang et al., 2006). This finding was later confirmed by binding experiments (Miao et al., 2007).

CONCLUSION

Protein structure prediction has been thought of as a "grand challenge" for some time now. As more and more researchers need and use the protein prediction tools, rapid progress has been made in recent years in this field. The massive amounts of sequence and structural data becoming available and the low cost and accessibility of computing power has led to an explosion of available tools and methods for protein prediction. The choice of one or another method still depends on the protein sequence, as well as the expected quality of the result. The rapid growth of automated servers means that protein prediction is no longer only for only a handful of researchers, but is available for the masses. The process in not completely automated,

the feedback of the user is still required when deciding on the most trustful method and the usefulness of the result.

REFERENCES

Anfinsen C.B. (1973). Principles that govern the folding of protein chains. *Science,* **181**(96), 223-230.

Arakaki Adrian K, Zhang Y., and *J. Skolnick* (2004). *Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment*. *Bioinformatics,* **20**(7), 1087-1096,

Arnold K., Bordoli L., Kopp J, and *T. Schwede* (2006a). The SWISS-MODEL workspace: a web-based environment for protein structure homology modeling. *Bioinformatics*, **22**(2), 195-201.

Bates P.A., Kelley L.A., MacCallum R.M., and *M.J.E. Sternberg* (2001). Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins-Structure Function and Genetics*, **5**(Suppl. 5), 39-46.

Benkert P., Tosatto C E., and *D. Schomburg* (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins,* **71**(1), 261-277.

Benson Dennis A., Karsch-Mizrachi I., Lipman J., Ostell J., and *W. Sayers* (2010). *GenBank*. *Nucleic Acids Res*. **38** (Database issue), D46-51.

Bonneau R., and *D. Baker (2001).* Ab initio protein structure prediction: progress and prospects. *Annu Rev Biophys Biomol Struct*. **30**, 173-189.

Bonneau R., Strauss EM., Rohl A., Chivian D., Bradley P., Malmström L., Robertson T., and *D. Baker (2002).* De Novo Prediction of Three-dimensional Structures for Major Protein Families. *Journal of Molecular Biology*, **322**(1), 65-78.

Bordoli L., Kiefer F., Arnold K., Benkert P., Battey J., and *T. Schwede* (2009a). Protein structure homology modeling using SWISS-MODEL workspace. *Nature Protocols,* **4**(1), 1-13.

Boyd A., Ciufo F., Barclay W., Graham E., Haynes P., Doherty K., Riesen M., Burgoyne D., and *A. Morgan* (2008). A random mutagenesis approach to isolate dominant-negative yeast sec1 mutants reveals a functional role for domain 3a in yeast and mammalian Sec1/Munc18 proteins. *Genetics*, **180**(1), 165-178.

Bujnicki, J. M. and *D. Fischer* (2004). 'Meta' approaches to protein structure prediction in J. M. Bujnicki (Ed.), *Nucleic Acids and Molecular Biology series: Practical Bioinformatics* (Vol. 23-24). Springer-Verlag.

Chivian D., and *D. Baker* (2006). Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic Acids Res,* **34**(17), e112.

Chivian D., Kim E., Malmström L., Bradley P., Robertson T., Murphy P., Strauss E M., Bonneau R., Rohl A., and *D. Baker* (2003). Automated prediction of CASP-5 structures using the Robetta server. *Proteins*, **53** Suppl 6, 524-533.

Combet C., Jambon M., Deleage G., and *C. Geourjon* (2002).: Geno3D: automatic comparative molecular modeling of protein. *Bioinformatics,* **18**(1), 213-214.

Coulson F W., and *J. Moult* (2002). A unifold, mesofold, and superfold model of protein fold use. *Proteins*, **46**(1), 61-71.

Du P., Andrec M., and *M. Levy* (2003). Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. *Protein Eng,* **16**(6), 407-414.

Dutta S., Burkhardt K., Young J., Swaminathan J., Matsuura T., Henrick K., Nakamura H., and *M. Berman* (2009). Data deposition and annotation at the worldwide protein data bank. *Mol Biotechnol*, **42**(1), 1-13.

Eisenberg D., Lüthy R., and *J U Bowie* (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol,* **277**(), 396-404.

Eswar N., Webb B., Marti-Renom A., Madhusudhan M S., Eramian D., Shen M., Pieper U., and A. Sali (2007). Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci*, Chapter 2, Unit 2.9.

Eyrich A., and *B. Rost* (2003). META-PP: single interface to crucial prediction servers. *Nucleic Acids Res*, **31**(13), 3308-3310.

Fischer D. (2006). Servers for protein structure prediction. *Curr Opin Struct Biol*, **16**(2), 178-182.

Fiser A., Do RKG., and *A. Sali* (2000). Modeling of loops in protein structures. *Protein Science,* **9**(9), 1753-1773.

Fiser A., and *A. Sali* (2003a). ModLoop: automated modeling of loops in protein structures. *Bioinformatics*, **19**(18), 2500-2501.

Fiser A. (2004). Protein structure modeling in the proteomics era. *Expert Rev Proteomics*, **1**(1), 97-110.

Fiser AS., and *A. Sali* (2003b). MODELLER: Generation and refinement of homology-based protein structure models. *Macromolecular crystallography*, 374(), 461+.

Floudas C A (2007). Computational methods in protein structure prediction. *Biotechnol Bioeng*, **97**(2), 207-213.

Floudas CA., Fung HK., McAllister SR., Mönnigmann M., and *R. Rajgaria* (2006). Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science*, **61**(3), 966-988.

Ginalski K., Elofsson A., Fischer D., and *L. Rychlewski* (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bionformatics*, **19**(8), 1015-1018.

Gough J., Karplus K., Hughey R., and *C. Chothia* (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology*, **313(**4), 903-919.

Gruenberg R., Nilges M., and *J. Leckner* (2007). Biskit - A software platform for structural bioinformatics. *Bioinformatics*, **23**(6), 769-770.

Hooft R W., Vriend G., Sander C., and *E E Abola* (1996). Errors in protein structures. *Nature*, **381**(6580), 272.

Howell P L., Almo S C., Parsons M R., Hajdu J., and G A Petsko (1992). Structure determination of turkey egg-white lysozyme using Laue diffraction data. *Acta Crystallogr B,* **48** (Pt 2)(), 200-207.

Jauch R., Yeo Hock C., Kolatkar R., and *D. Clarke* (2007). Assessment of CASP7 structure predictions for template free targets. *Proteins*, **69** Suppl 8(), 57-67.

Jun-Tao, G., Kyle, E. and *X. Ying* (2008). A Historical Perspective of Template-Based Protein Structure Prediction in Z. Mohammed and B. Christopher (Eds.), *Protein Structure Prediction* (Vol. 4, 3-42). Humana Press.

Kaufmann W., Lemmon H., DeLuca L., Sheehan H., and *J. Meiler* (2010). Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. *Biochemistry*, **49**(14), 2987-2998.

Kelley LA., MacCallum RM., and *MJE Sternberg* (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *Journal of Molecular Biology*, **299**(2), 499-520.

Kelley A., and *J E Sternberg* (2009a). Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc*, **4**(3), 363-371.

Kiefer F., Arnold K., Kuenzli M., Bordoli L., and *T. Schwede* (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Research*, **37**(Sp. Iss. SI), D387-D392.

Kihara D., Chen H., and *D. Yang* (2009). Quality assessment of protein structure models. *Curr Protein Pept Sci,* **10**(3), 216-228.

Kihara Daisuke., and *J. Skolnick* (2003). The PDB is a Covering Set of Small Protein Structures. *Journal of Molecular Biology*, **334**(4), 793-802.

Kim E., Chivian D., and *D. Baker* (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res*, **32**(Web Server issue), W526-31.

Koh Y Y., Eyrich A., Marti-Renom A., Przybylski D., Madhusudhan S., Eswar N., Graña O., Pazos F., Valencia A., Sali A., and *B. Rost* (2003). *EVA: Evaluation of protein structure prediction servers. Nucleic Acids Res*, **31**(13), 3311-3315.

Kolata G. (1986). Trying to crack the second half of the genetic code. *Science*, **233**(4768), 1037-1039.

Kolinski A. (2004). Protein modeling and structure prediction with a reduced representation. *Acta Biochimica Polonica*, **51**(2), 349-371.

Kopp J., Bordoli L., Battey N D., Kiefer F., and *T. Schwede* (2007). Assessment of CASP7 predictions for template-based modeling targets. *Proteins*, **69** Suppl 8, 38-56.

Kopp J., and *T. Schwede* (2004). Automated protein structure homology modeling: a progress report. *Pharmacogenomics*, **5**(4), 405-416.

Kryshtafovych A., Fidelis K., and *J. Moult* (2009). CASP8 results in context of previous experiments. *Proteins*, **77** Suppl 9(), 217-228.

Kurowski MA., and *J.M. Bujnicki* (2003). GeneSilico protein structure prediction meta-server. *Nucleic Acids Research*, **31**(13), 3305-3307.

Lambert C., Leonard N., De Bolle X., and *E. Depiereux* (2002): ESyPred3D: Prediction of proteins 3D structures. *Bioinformatics*, **18**(9), 1250-1256.

Laskowski RA., Macarthur MW., Moss DS., and *JM Thornton* (1993). {PROCHECK}: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst*, **26**(), 283-291.

Lattman E. (1995). Protein structure prediction: A special issue. *Proteins: Structure, Function, and Genetics*, **23**(3), i.

Lesk A M., and *C. Chothia* (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J *Mol Biol,* **136**(3), 225-270.

Madera M., Vogel C., Kummerfeld SK., Chothia C., and *J. Gough* (2004). The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Research*, **32**(Sp. Iss. SI), D235-D239.

Meller J., and *R. Elber* (2001). Linear programming optimization and a double statistical filter for protein threading protocols. *Proteins: Structure, Function, and Genetics*, **45**(3), 241-261.

Miao Zhenhua., Luker E., Summers C., Berahovich R., Bhojani S., Rehemtulla A., Kleer G., Essner J., Nasevicius A., Luker D., Howard C., and *J. Schall* (2007). CXCR7 (RDC1) promotes breast and lung tumor growth in vivo and is expressed on tumor-associated vasculature. *Proc Natl Acad Sci U S A*, **104**(40), 15735-15740.

Modi S., Paine M J., Sutcliffe M J., Lian L Y., Primrose W U., Wolf C R., and *G C Roberts* (1996). A model for human cytochrome P450 2D6 based on homology modeling and NMR studies of substrate binding. *Biochemistry*, **35**(14), 4540-4550,

Peitsch MC. (1995). Protein modeling by e-mail. *Bio-Technology*, **13**(7), 658-660.

Petrey D., Xiang Z., Tang L., Xie L., Gimpelev M., Mitros T., Soto S., Goldsmith-Fischman S., Kernytsky A., Schlessinger A., Koh Y Y., Alexov E., and *B. Honig* (2003). Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*, **53** Suppl 6, 430-435.

Pieper U., Eswar N., Stuart C., Ilyin A., and *A. Sali* (2002). MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res*, **30**(1), 255-259.

Pillardy J., Czaplewski C., Liwo A., Lee J., Ripoll R., Kaźmierkiewicz R., Ołdziej S., Wedemeyer J., Gibson D., Arnautova A., Saunders J., Ye Y., and *A. Scheraga* (2001). Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(5), 2329-2333.

Read J., and *G. Chavali* (2007). Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins*, **69** Suppl 8, 27-37.

Ring C S., Sun E., McKerrow J H., Lee G K., Rosenthal P J., Kuntz I D., and *F E Cohen* (1993). Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc Natl Acad Sci U S A*, **90**(8), 3583-3587.

Rost B, (1996). PHD: Predicting one-dimensional protein structure by profile-based neural networks. *Computer methods for macromolecular sequence analysis,* **266**(), 525-539.

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

Rychlewski L., and *D. Fischer* (2005). LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci*, **14**(1), 240-245.

Sali A., and *T L Blundell* (1993). Comparative protein modeling by satisfaction of spatial restraints. *J Mol Biol*, **234**(3), 779-815.

Sali A., Potterton L., Yuan F., Vanvlijmen H., and *M. Karplus* (1995). Evaluation of comparative protein modeling by MODELER. *Proteins: Structure, Function, and Genetics*, **23**(3), 318-326.

Sanchez R., and *A. Sali* (1997). Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins: Structure, Function, and Genetics*, (Suppl. 1), 50-58.

Schymkowitz J., Borg J., Stricher F., Nys R., Rousseau F., and *L. Serrano* (2005). The FoldX web server: an online force field. *Nucleic Acids Research*, **33** (Suppl. 2), W382-W388.

Sheng Y., Sali A., Herzog H., Lahnstein J., and *S A Krilis* (1996): Site-directed mutagenesis of recombinant human beta 2-glycoprotein I identifies a cluster of lysine residues that are critical for phospholipid binding and anti-cardiolipin antibody activity. *J Immunol*, **157**(8), 3744-3751.

Simons T., Kooperberg C., Huang E., and *D. Baker* (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, **268** (1), 209-225.

Sippl M J. (1995). Knowledge-based potentials for proteins. *Curr Opin Struct Biol*, **5**(2), 229-235.

Soding J., Biegert A., and *AN Lupas* (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research,* **33** (Suppl. 2), W244-W248.

Teodorescu O., Galor T., Pillardy J., and *R. Elber* (2004). Enriching the sequence substitution matrix by structural information. *Proteins: Structure, Function, and Bioinformatics*, **54** (1), 41-48.

Tobi D., and *R. Elber (2000)*. Distance-dependent, pair potential for protein folding: Results from linear optimization. *Proteins: Structure, Function, and Genetics*, **41** (1), 40-46.

Tress M., Cheng J., Baldi P., Joo K., Lee J., Seo J., Lee J., Baker D., Chivian D., Kim D., and *I. Ezkurdia* (2007). Assessment of predictions submitted for the CASP7 domain prediction category. *Proteins*, **69** Suppl 8, 137-151.

Vakser I A. (1995). Protein docking for low-resolution structures. *Protein Eng*, **8**(4), 371-377.

Vernal J., Fiser A., Sali A., Müller M., Cazzulo J., and *C. Nowicki* (2002). Probing the specificity of a trypanosomal aromatic alpha-hydroxy acid dehydrogenase by site-directed mutagenesis. *Biochem Biophys Res Commun*, **293**(1), 633-639.

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

Vriend G. (1990). WHAT IF - a molecular modeling and drug design program. *Journal of Molecular Graphics*, 8(1), 52.

Wallner B., and *A. Elofsson* (2005). All are not equal: a benchmark of different homology modeling programs. *Protein Sci*, 14(5), 1315-1327.

Wilson D., Madera M., Vogel C., Chothia C., and *J. Gough* (2007). The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Research,* **35**(Sp. Iss. SI), D308-D313.

Wilson D., Pethica R., Zhou Y., Talbot C., Vogel C., Madera M., Chothia C., and *J. Gough* (2009). SUPERFAMILY-sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Research* 37(Sp. Iss. SI), D380-D386.

Wu G., Fiser A., ter Kuile B., Sali A., and *M. Müller* (1999). Convergent evolution of Trichomonas vaginalis lactate dehydrogenase from malate dehydrogenase. *Proc Natl Acad Sci U S A*, **96**(11), 6285-6290.

Wu S., and *Y. Zhang (2007)*. LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research*, **35**(10), 3375-3382.

Wu S., and *Y Zhang* (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins: Structure, Function, and Bioinformatics*, **72**(2), 547-556.

Xiang Z. (2006). Advances in homology protein structure modeling. *Curr Protein Pept Sci*, **7**(3), 217-227.

Zhang Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins: Structure, Function, and Bioinformatics*, **69**(Suppl. 8), 108-117.

Zhang Y. (2008a). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, **9**(Suppl. 9), 100-113.

Zhang Y. (2008b). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, **18**(3), 342-348.

Zhang Y. (2009a). I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins: Structure, Function, and Bioinformatics*, **77**(Suppl. 9), 100-113.

Zhang Y. (2009b). Protein structure prediction: when is it useful?. *Current Opinion in Structural Biology*, **19**(2), 145-155.

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

Zhang Y., and *J. Skolnick* (2005). The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(4), 1029-1034.