Preface

Bioinformatics is the management and analysis of data for the life sciences. As such, it is inherently interdisciplinary, drawing on techniques from Computer Science, Statistics, and Mathematics and bringing them to bear on problems in Biology. Moreover, its subject matter is as broad as Biology itself. Users and developers of bioinformatics methods come from all of these fields. Molecular biologists are some of the major users of Bioinformatics, but its techniques are applicable across a range of life sciences. Other users include geneticists, microbiologists, biochemists, plant and agricultural scientists, medical researchers, and evolution researchers.

The ongoing exponential expansion of data for the life sciences is both the major challenge and the *raison d'être* for twenty-first century Bioinformatics. To give one example among many, the completion and success of the human genome sequencing project, far from being the end of the sequencing era, motivated a proliferation of new sequencing projects. And it is not only the quantity of data that is expanding; new types of biological data continue to be introduced as a result of technological development and a growing understanding of biological systems.

Bioinformatics describes a selection of methods from across this vast and expanding discipline. The methods are some of the most useful and widely applicable in the field. Most users and developers of bioinformatics methods will find something of value to their own specialties here, and will benefit from the knowledge and experience of its 86 contributing authors. Developers will find them useful as components of larger methods, and as sources of inspiration for new methods. Volume II, Section IV in particular is aimed at developers; it describes some of the "meta-methods"—widely applicable mathematical and computational methods that inform and lie behind other more specialized methods—that have been successfully used by bioinformaticians. For users of bioinformatics, this book provides methods that can be applied as is, or with minor variations to many specific problems. The Notes section in each chapter provides valuable insights into important variations and when to use them. It also discusses problems that can arise and how to fix them. This work is also intended to serve as an entry point for those who are just beginning to discover and use methods in bioinformatics. As such, this book is also intended for students and early career researchers.

As with other volumes in the Methods in Molecular Biology™ series, the intention of this book is to provide the kind of detailed description and implementation advice that is crucial for getting optimal results out of any given method, yet which often is not incorporated into journal publications. Thus, this series provides a forum for the communication of accumulated practical experience.

The work is divided into two volumes, with data, sequence analysis, and evolution the subjects of the first volume, and structure, function, and application the subjects of the second. The second volume also presents a number of "meta-methods": techniques that will be of particular interest to developers of bioinformatic methods and tools.

Within Volume I, Section I deals with data and databases. It contains chapters on a selection of methods involving the generation and organization of data, including

sequence data, RNA and protein structures, microarray expression data, and functional annotations.

Section II presents a selection of methods in sequence analysis, beginning with multiple sequence alignment. Most of the chapters in this section deal with methods for discovering the functional components of genomes, whether genes, alternative splice sites, non-coding RNAs, or regulatory motifs.

Section III presents several of the most useful and interesting methods in phylogenetics and evolution. The wide variety of topics treated in this section is indicative of the breadth of evolution research. It includes chapters on some of the most basic issues in phylogenetics: modelling of evolution and inferring trees. It also includes chapters on drawing inferences about various kinds of ancestral states, systems, and events, including gene order, recombination events and genome rearrangements, ancestral interaction networks, lateral gene transfers, and patterns of migration. It concludes with a chapter discussing some of the achievements and challenges of algorithm development in phylogenetics.

In Volume II, Section I, some methods pertinent to the prediction of protein and RNA structures are presented. Methods for the analysis and classification of structures are also discussed.

Methods for inferring the function of previously identified genomic elements (chiefly protein-coding genes) are presented in Volume II, Section II. This is another very diverse subject area, and the variety of methods presented reflects this. Some well-known techniques for identifying function, based on homology, "Rosetta stone" genes, gene neighbors, phylogenetic profiling, and phylogenetic shadowing are discussed, alongside methods for identifying regulatory sequences, patterns of expression, and participation in complexes. The section concludes with a discussion of a technique for integrating multiple data types to increase the confidence with which functional predictions can be made. This section, taken as a whole, highlights the opportunities for development in the area of functional inference.

Some medical applications, chiefly diagnostics and drug discovery, are described in Volume II, Section III. The importance of microarray expression data as a diagnostic tool is a theme of this section, as is the danger of over-interpreting such data. The case study presented in the final chapter highlights the need for computational diagnostics to be biologically informed.

The final section presents just a few of the "meta-methods" that developers of bioinformatics methods have found useful. For the purpose of designing algorithms, it is as important for bioinformaticians to be aware of the concept of *fixed parameter tractability* as it is for them to understand NP-completeness, since these concepts often determine the types of algorithms appropriate to a particular problem. *Clustering* is a ubiquitous problem in Bioinformatics, as is the need to *visualize* data. The need to interact with massive data bases and multiple software entities makes the development of *computational pipelines* an important issue for many bioinformaticians. Finally, the chapter on *text mining* discusses techniques for addressing the special problems of interacting with and extracting information from the vast biological literature.

Jonathan M. Keith

Chapter 2

Protein Structure Prediction

Bissan Al-Lazikani, Emma E. Hill, and Veronica Morea

Abstract

Protein structure prediction has matured over the past few years to the point that even fully automated methods can provide reasonably accurate three-dimensional models of protein structures. However, until now it has not been possible to develop programs able to perform as well as human experts, who are still capable of systematically producing better models than automated servers. Although the precise details of protein structure prediction procedures are different for virtually every protein, this chapter describes a generic procedure to obtain a three-dimensional protein model starting from the amino acid sequence. This procedure takes advantage both of programs and servers that have been shown to perform best in blind tests and of the current knowledge about evolutionary relationships between proteins, gained from detailed analyses of protein sequence, structure, and functional data.

Key words: Protein structure prediction, homology modeling, fold recognition, fragment assembly, metaservers.

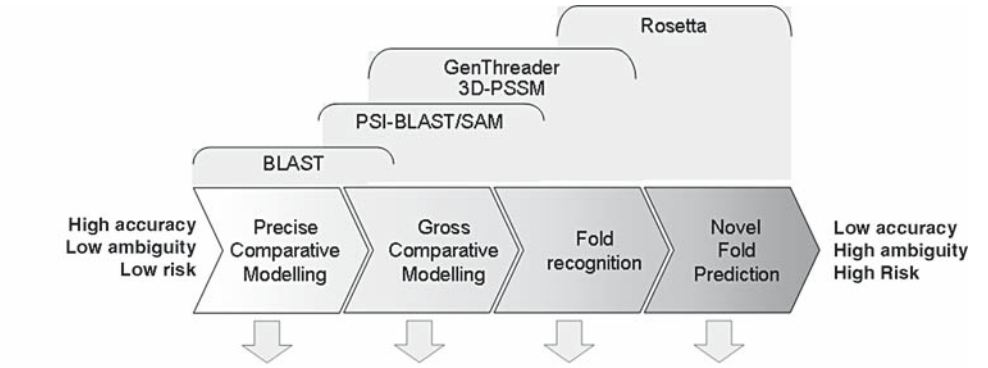
1. Introduction

In spite of many years of intense research, unravelling the algorithm by which Nature folds each amino acid (a.a.) sequence into a unique protein three-dimensional (3D) structure remains one of the great unsolved problems in molecular biology. However, analyses of the wealth of information contained in protein sequence and structural databases (DBs) have revealed the existence of a number of fundamental rules and relationships among protein sequence, structure, and function, based on which many of both the current theories about molecular evolution and protein structure prediction methods have been developed.

The first important question to ask when dealing with protein structure prediction concerns the purpose for which the

Jonathan M. Keith (ed.), *Bioinformatics, Volume II: Structure, Function and Applications, vol. 453*

Book doi: 10.1007/978-1-60327-429-6 Springerprotocols.com



Drug Design, Docking	+++			
SAR rationalization	+++	++		
Construct design	+++	+++		
Function prediction	+++	+++	++	+
Assay selection	+++	++	+	

Fig. 2.1. Biological applications of protein structure prediction methods. SAR: structure–activity relationships.

model is built. This is of fundamental importance since the accuracy required of the model, that is, its similarity to the real protein structure, can be substantially different for different biological applications (**Fig. 2.1**). At one end of the spectrum, a very accurate prediction of the functional site in terms of both main- and side-chain conformation is indispensable for drug design purposes, and a correct positioning of the side-chains involved in intermolecular interactions is required for protein–protein interaction (docking) studies. At the other extreme, an approximate description of the protein topology at the level of general arrangement of secondary structure elements (SSE) or domains, or even an idea of which regions are likely to be globular, unfolded, or aggregation prone, can be valuable to those who want to cut insoluble proteins into smaller and soluble portions, which are likely to be more easily expressed and studied experimentally. In general, correct models of the overall protein fold, even with unrefined details, can be useful to rationalize experimental data at a structural level and guide the design of new experiments aimed at improving our understanding of protein function. In the era of structural genomics and fast solving of protein structures, looking for structural biologists interested in experimentally determining the structure of your protein(s) is also an option.

In choosing the procedure to follow for model building, other factors to consider are the time available and number of models to make. Production of 3D models on a large and even genomic scale is achievable with the use of automated or partially automated methods. A number of automated methods have been developed

that can rapidly produce 3D models starting from the amino acid sequence of a target protein (*see* **Section 3.5.1**). However, the quality of these models can vary to a large extent. A detailed understanding of the underlying steps of the modeling procedure is required to evaluate, and often improve, the accuracy of automatically produced models.

The most reliable source of information about the accuracy of protein structure prediction methods is provided by the evaluation of their performance in blind tests. In such evaluations, 3D models of target proteins are compared with the experimentally determined structures of the same proteins using visual assessments performed by human experts and/or numerical evaluators of structural similarity (**Note 1**). Two main types of evaluations are performed on a regular basis: fully automated evaluations (devoted to fully automated methods), and human-based evaluations (examining predictions refined by human experts as well as those provided by fully automated methods). The human-based evaluation, named Critical Assessment of Structure Predictions (*CASP*), is performed every 2 years since its debut in 1994 and contributes enormously to the improvement of protein structure prediction methods as well as to the diffusion of information about their performance. (The URLs of all the Web sites, programs, servers, and databases indicated in italic in the text are reported in **Table 2.1**, along with relevant references, when available.) A full description of the experiment along with reports of the results of each of the six previous *CASP* experiments are available *(1–6)*. A seventh edition took place in 2006 and its results, published in 2007, provides up-to-date information on the most recent advances in the field of protein structure prediction (preliminary evaluation results are available from the *CASP7* Web site). In parallel with the last four *CASP* editions, the Critical Assessment of Fully Automated Structure Predictions (*CAFASP*) experiments have also been run in which the ability of automated servers to predict *CASP* targets was evaluated by other servers, without human intervention *(7–9)*. In both *CASP* and *CAFASP* the predictions are guaranteed to be "blind" by the fact that they are submitted to the evaluation before the experimental structures are released. The performance of automated servers on proteins recently released from the *PDB (10)* is also evaluated on a continuous basis by servers such as *Livebench (11)* and *EVA (12)*. Every week these servers submit the sequences of proteins newly released from the *PDB* to the prediction servers participating in these experiments, collect their results, and evaluate them using automated programs. To take part in *Livebench* and *EVA* the prediction servers must agree to delay the updating of their structural template libraries by 1 week, as the predictions evaluated in these experiments refer to targets whose structures have already been made publicly available. These predictions cannot, therefore, be considered strictly "blind" (as those evaluated in *CASP* and *CAFASP*). Nevertheless, automated assessments can provide an ongoing picture of how automated prediction methods

Table 2.1
URLs of web sites, programs, servers and DBs relevant to protein structure prediction

Web site, program, server, DB	URL	References
@TOME	bioserv.cbs.cnrs.fr/HTML_BIO/frame_meta.html	*(210)*
3D-JIGSAW	www.bmm.icnet.uk/servers/3djigsaw/	*(164)*
3D-Jury	BioInfo.PL/Meta/	*(196)*
3D-PSSM	www.sbg.bio.ic.ac.uk/~3dpssm/	*(94)*
3D-Shotgun	www.cs.bgu.ac.il/~bioinbgu/	*(197–200)*
ANOLEA	protein.bio.puc.cl/cardex/servers/anolea/index.html	*(173, 174)*
Arby	arby.bioinf.mpi-inf.mpg.de/arby/jsp/index.jsp	*(80, 81)*
BCM Search Launcher	searchlauncher.bcm.tmc.edu/	*(34)*
Belvu	www.cgb.ki.se/cgb/groups/sonnhammer/Belvu.html	*(110)*
BioEdit	www.mbio.ncsu.edu/BioEdit/bioedit.html	
Bioinbgu	www.cs.bgu.ac.il/~bioinbgu/	*(197–200)*
BLAST	www.ncbi.nlm.nih.gov/BLAST/[a]	*(67)*
CAFASP4 MQAPs	cafasp4.cse.buffalo.edu/progs/mqaps/	
CAFASP	www.cs.bgu.ac.il/~dfischer/CAFASP5/	*(7–9)*
CAPRI	capri.ebi.ac.uk/	*(36, 37)*
CASP experiments (CASP1-CASP7)	predictioncenter.org/	*(1–6)*
CATH	www.biochem.ucl.ac.uk/bsm/cath/cath.html	*(29)*
CAZy	afmb.cnrs-mrs.fr/CAZY/	*(25)*
CDD	www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi	*(41)*
CE	cl.sdsc.edu/	*(122)*
CINEMA	umber.sbs.man.ac.uk/dbbrowser/CINEMA2.1/	*(112)*
ClustalW	www.ebi.ac.uk/clustalw/	*(106)*
ClustalX	ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/	*(108)*
Cn3D	www.ncbi.nlm.nih.gov/Structure/CN3D/cn3dinstall.shtml	*(147)*
COACH	www.drive5.com/lobster/	*(86)*
COLORADO3D	asia.genesilico.pl/colorado3d/	*(142)*
COMPASS	prodata.swmed.edu/compass/compass.php	*(82–84)*
CONSURF	consurf.tau.ac.il/	*(143)*

(continued)

Table 2.1 (continued)

Web site, program, server, DB	URL	References
CPHmodels	www.cbs.dtu.dk/services/CPHmodels/	*(168)*
DALI	www.ebi.ac.uk/dali/	*(31)*
DaliLite	www.ebi.ac.uk/DaliLite/	*(119)*
DISOPRED	bioinf.cs.ucl.ac.uk/disopred/	*(55)*
DISpro	www.ics.uci.edu/~baldig/dispro.html	*(52)*
Domain Parser	compbio.ornl.gov/structure/domainparser/	*(114, 115)*
DomCut	www.bork.embl.de/~suyama/domcut/	*(40)*
DOMPLOT	www.biochem.ucl.ac.uk/bsm/domplot/index.html	*(117)*
DRIPPRED	www.sbc.su.se/~maccallr/disorder/	
DSSP	swift.cmbi.ru.nl/gv/dssp/	*(92)*
EBI	www.ebi.ac.uk	
Entrez Tutorial	www.ncbi.nlm.nih.gov/Entrez/tutor.html	
ESyPred3D	www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/	*(169)*
EVA	eva.compbio.ucsf.edu/~eva/	*(12)*
Expasy	www.expasy.org	*(33)*
FAMSBASE	daisy.nagahama-i-bio.ac.jp/Famsbase/index.html	*(214)*
FastA	www.ebi.ac.uk/fasta33/	*(69)*
Fasta format	www.ebi.ac.uk/help/formats_frame.html	
FFAS03	ffas.burnham.org	*(88)*
FORTE	www.cbrc.jp/forte	*(89)*
FRankenstein3D	genesilico.pl/frankenstein	*(207)*
FSSP	ekhidna.biocenter.helsinki.fi/dali/start	*(30, 31)*
Fugue	www-cryst.bioc.cam.ac.uk/fugue/	*(99)*
Genesilico	www.genesilico.pl/meta/	*(202)*
GenThreader, mGen-Threader	bioinf.cs.ucl.ac.uk/psipred/psiform.html	*(95, 96)*
Ginzu	robetta.bakerlab.org/	*(39)*
GROMACS	www.gromacs.org/	*(179)*
HBPLUS	www.biochem.ucl.ac.uk/bsm/hbplus/home.html	*(137)*
HHpred	toolkit.tuebingen.mpg.de/hhpred	*(91)*
HMAP	trantor.bioc.columbia.edu/hmap/	*(101)*

(continued)

Table 2.1 (continued)

Web site, program, server, DB	URL	References
HMMER	hmmer.wustl.edu/	*(72, 73)*
Homo sapiens genome	www.ensembl.org/Homo_sapiens/index.html	*(26)*
Homstrad	www-cryst.bioc.cam.ac.uk/~homstrad/	*(32)*
IMPALA	blocks.fhcrc.org/blocks/impala.html	*(77)*
InsightII / Biopolymer / Discover	www.accelrys.com/products/insight/	*(178)*
InterPro	www.ebi.ac.uk/interpro/	*(45)*
Inub	inub.cse.buffalo.edu/	*(199, 200)*
IUPred	iupred.enzim.hu/index.html	*(53)*
Jackal	trantor.bioc.columbia.edu/programs/jackal/index.html	
Joy	www-cryst.bioc.cam.ac.uk/joy/	*(135)*
Jpred	www.compbio.dundee.ac.uk/~www-jpred/	*(65, 66)*
LAMA	blocks.fhcrc.org/blocks-bin/LAMA_search.sh	*(87)*
LGA	predictioncenter.org/local/lga/lga.html	*(118)*
LIGPLOT	www.biochem.ucl.ac.uk/bsm/ligplot/ligplot.html	*(140)*
Livebench	bioinfo.pl/meta/livebench.pl	*(11)*
LOBO	protein.cribi.unipd.it/lobo/	*(156)*
LOOPP	cbsuapps.tc.cornell.edu/loopp.aspx	*(102)*
Loopy	wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:Loopy	*(155)*
Mammoth	ub.cbm.uam.es/mammoth/pair/index3.php	*(120)*
Mammoth-mult	ub.cbm.uam.es/mammoth/mult/	*(126)*
Meta-BASIC	BioInfo.PL/Meta/	*(113)*
ModBase	modbase.compbio.ucsf.edu/modbase-cgi-new/search_form.cgi	*(213)*
Modeller	salilab.org/modeller/	*(161)*
ModLoop	alto.compbio.ucsf.edu/modloop/	*(157)*
MolMol	hugin.ethz.ch/wuthrich/software/molmol/index.html	*(149)*
MQAP-Consensus	cafasp4.cse.buffalo.edu/mqap/submit.php	*(177)*
NACCESS	wolf.bms.umist.ac.uk/naccess/	*(136)*
NAMD	www.ks.uiuc.edu/Research/namd/	*(180)*
NCBI	www.ncbi.nlm.nih.gov	
NCBI NR sequence DB	ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz	*(23)*

(continued)

Table 2.1 (continued)

Web site, program, server, DB	URL	References
Nest	wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:nest	*(165)*
ORFeus	bioinfo.pl/meta/	*(90)*
Pcons, Pmodeller	www.bioinfo.se/pcons/, www.bioinfo.se/pmodeller/	*(201)*
Pcons5	www.sbc.su.se/~bjornw/Pcons5/	*(208, 209)*
PDB	www.pdb.org/	*(10)*
PDBsum	www.ebi.ac.uk/thornton-srv/databases/pdbsum/	*(141)*
PDP	123d.ncifcrf.gov/pdp.html	*(116)*
Pfam	www.sanger.ac.uk/Software/Pfam/	*(42)*
Phyre	www.sbg.bio.ic.ac.uk/~phyre/	
Picasso	www.embl-ebi.ac.uk/picasso/	*(85)*
PMDB	a.caspur.it/PMDB/	*(215)*
Porter	distill.ucd.ie/porter/	*(64)*
POSA	fatcat.burnham.org/POSA/	*(127)*
PPRODO	gene.kias.re.kr/~jlee/pprodo/	*(40)*
PRC	supfam.mrc-lmb.cam.ac.uk/PRC/	
PRED-TMBB	Biophysics.biol.uoa.gr/PRED-TMBB/	*(60)*
PredictProtein	www.predictprotein.org/	*(59)*
PrISM	wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:PrISM	*(125)*
Procheck	www.biochem.ucl.ac.uk/~roman/procheck/procheck.html	*(138, 139)*
ProDom	protein.toulouse.inra.fr/prodom/current/html/home.php	*(44)*
PROF	cubic.bioc.columbia.edu/predictprotein/	*(59)*
ProQ, ProQres	www.sbc.su.se/~bjornw/ProQ/	*(175, 176)*
ProSa	www.came.sbg.ac.at/typo3/	*(171)*
Prosite	www.expasy.org/prosite/	*(46)*
Protein Explorer	proteinexplorer.org	*(146)*
Protinfo AB CM	protinfo.compbio.washington.edu/protinfo_abcmfr/	*(159)*
PSI-BLAST	www.ncbi.nlm.nih.gov/BLAST/	*(67)*
Psi-Pred	bioinf.cs.ucl.ac.uk/psipred/	*(57, 62)*
RAPTOR	ttic.uchicago.edu/~jinbo/RAPTOR_form.htm	*(100)*
RasMol	www.umass.edu/microbio/rasmol/getras.htm	*(145)*

(continued)

Table 2.1 (continued)

Web site, program, server, DB	URL	References
ReadSeq	bioweb.pasteur.fr/seqanal/interfaces/readseq-simple.html	
Robetta	robetta.bakerlab.org/	*(187–189)*
ROKKY	www.proteinsilico.org/rokky/	*(194)*
Rosetta	depts.washington.edu/ventures/UW_Technology/Express_Licenses/Rosetta/	*(19, 182–186)*
Rosettadom	robetta.bakerlab.org/	*(39)*
SAM (download)	www.soe.ucsc.edu/research/compbio/sam2src/	
SAM-T02	www.cse.ucsc.edu/research/compbio/HMM-apps/T02-query.html	*(75, 76)*
SAM-T99	www.cse.ucsc.edu/research/compbio/HMM-apps/T99-query.html	*(63)*
Sanger Centre	www.sanger.ac.uk	
Scap	wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:Scap	*(160)*
Schistosoma mansoni genome	www.tigr.org/tdb/e2k1/sma1/	*(27)*
SCOP	scop.mrc-lmb.cam.ac.uk/scop/	*(28)*
SCRWL	www1.jcsg.org/scripts/prod/scwrl/serve.cgi	*(158)*
Seaview	pbil.univ-lyon1.fr/software/seaview.html	*(111)*
SegMod/ENCAD	csb.stanford.edu/levitt/segmod/	*(162)*
Sequence Manipulation Suite	bioinformatics.org/sms2/	*(35)*
SMART	smart.embl-heidelberg.de/	*(43)*
SP3	sparks.informatics.iupui.edu/hzhou/anonymous-fold-sp3.html	*(104, 105)*
SPARKS2	sparks.informatics.iupui.edu/hzhou/anonymous-fold-sparks2.html	*(103, 104)*
SPRITZ	protein.cribi.unipd.it/spritz/	*(54)*
SSAP	www.cathdb.info/cgi-bin/cath/GetSsapRasmol.pl	*(123)*
SSEARCH	pir.georgetown.edu/pirwww/search/pairwise.shtml	*(70)*
SSM	www.ebi.ac.uk/msd-srv/ssm/ssmstart.html	*(124)*
STRUCTFAST	www.eidogen-sertanty.com/products_tip_structfast.html	*(195)*
SUPERFAMILY	supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/	*(47, 48)*
Swiss-PDBViewer	www.expasy.org/spdbv/[b]	*(144)*

(continued)

Table 2.1 (continued)

Web site, program, server, DB	URL	References
SwissModel	swissmodel.expasy.org/	*(163)*
SwissModel Repository	swissmodel.expasy.org/repository/	*(212)*
SwissProt, TrEMBL	www.expasy.uniprot.org/database/download.shtml	
T-Coffee	igs-server.cnrs-mrs.fr/~cnotred/Projects_home_page/t_coffee_home_page.html	*(107)*
TASSER-Lite	cssb.biology.gatech.edu/skolnick/webservice/tasserlite/index.html	*(97)*
TBBpred	www.imtech.res.in/raghava/tbbpred/	*(61)*
Threader	bioinf.cs.ucl.ac.uk/threader/	*(16)*
Three to One	bioinformatics.org/sms2/three_to_one.html	
TIGR	www.tigr.org	
TINKER	dasher.wustl.edu/tinker/	
TMHMM	www.cbs.dtu.dk/services/TMHMM/	*(58)*
Translate	www.expasy.org/tools/dna.html	
UniProt	www.expasy.uniprot.org/	*(24)*
VAST	www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html	*(121)*
Verify-3D	nihserver.mbi.ucla.edu/Verify_3D/	*(15, 172)*
VMD	www.ks.uiuc.edu/Research/vmd/	*(150)*
VSL2	www.ist.temple.edu/disprot/predictorVSL2.php	*(50)*
WebLogo	weblogo.berkeley.edu/logo.cgi	*(109)*
Whatcheck	www.cmbi.kun.nl/gv/whatcheck/	*(170)*
WHAT IF	swift.cmbi.kun.nl/whatif/	*(148)*
Wikipedia on Structural Alignment Software	en.wikipedia.org/wiki/Structural_alignment_software#Structural_alignment	

[a]For more details, see the BLAST tutorial (www.ncbi.nlm.nih.gov/BLAST/tutorial/) and frequently asked questions (FAQ) (www.ncbi.nlm.nih.gov/blast/blast_FAQs.shtml).
[b]Also download: Swiss-Pdb Viewer Loop Database, User Guide, and Tutorial, containing detailed information on the program commands and explanations on how to build a homology model of the target protein using this program.

perform based on a larger number of targets than those evaluated by *CASP/CAFASP* experiments.

In *CASP* and related experiments protein structure prediction methods have been traditionally grouped into three broad categories depending on the level of similarity of the target protein sequence to other proteins of known structure, which necessarily impacts the

procedure that must be used to build the models: comparative or homology modeling (CM), fold recognition (FR), and new fold (NF) predictions. The separation between these categories, in particular between CM and FR and between FR and NF, has been challenged by the development of more sophisticated methods (e.g., profile-based methods and fragment-based methods, *see* **Sections 3.3.1.1** and **3.4**) able to cross the boundaries between them. The accuracy of the structures predicted in blind tests is generally highest for CM and lowest for NF methods, but there is a large overlap between the accuracy reached by neighboring categories of methods (**Fig. 2.2**).

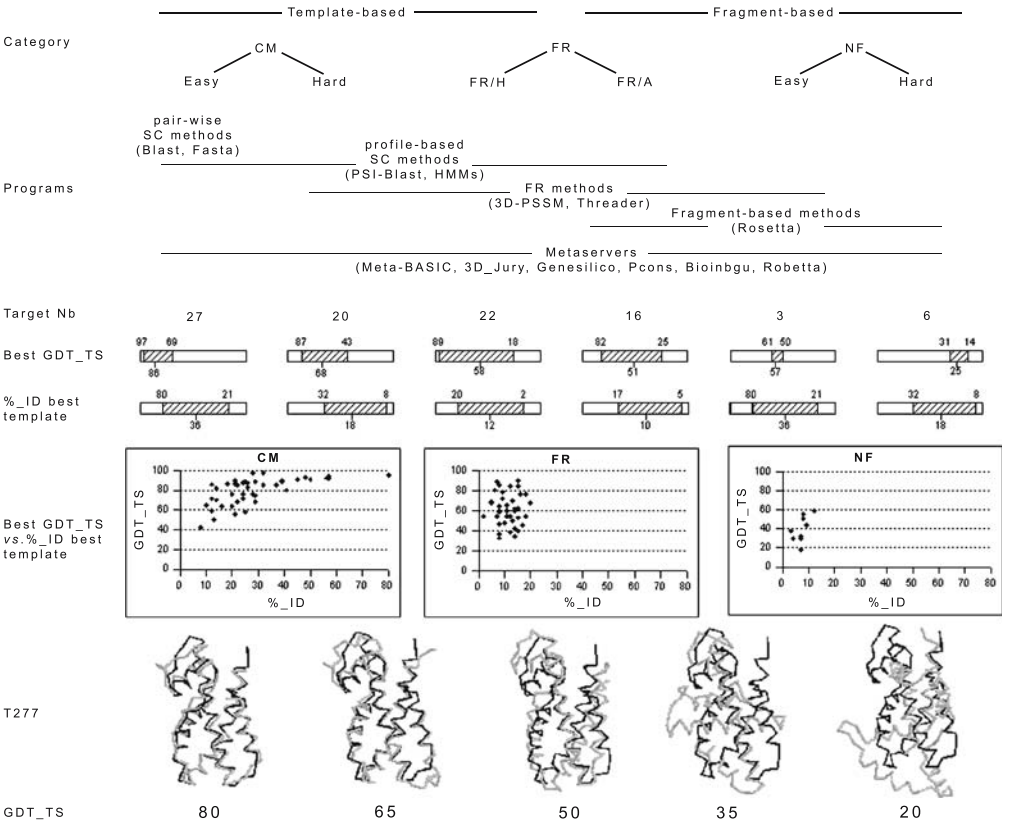


Fig. 2.2. Protein structure prediction methods used in *CASP6* and accuracy of the predictions, expressed by GDT_TS (Note 1). For a description of prediction methods and *CASP6* prediction categories see text. Target Nb, Best GDT_TS, and %_ID best template are the number of targets evaluated in each prediction category, the GDT_TS between the real structure of each target and the best model submitted for it, and the %_ID between the target sequence and the best template structure present in the *PDB*, i.e., the structure most similar to the target. The bottom panel shows the structural superposition between the experimental structure of *CASP6* target T277 (PDB ID: 1wty) and five models with varying degrees of accuracy submitted by predictors taking part in the experiment. Only Cα atoms of structure *(black)* and models *(gray)* are shown. As shown by the figure, GDT_TS values near 80 indicate that the Cα atoms of the model are well superimposed to those of the structure, except in some loop regions. For a GDT_TS value of 65 the core regions are still predicted quite accurately, although structural differences in the loop regions and protein *termini* are more pronounced. GDT_TS values around 50 correspond to an overall structural similarity, but structural variations occur even in the conserved core regions. GDT_TS values around 35 indicate lack of an overall accurate topology prediction, with similarity between only about half of the core regions of structure and model, whereas other regions differ significantly. For GDT_TS values around 20, only a small fraction of the model shows some resemblance to the target structure.

The rational basis of Comparative Modeling (CM) are the following observations: (1) proteins having highly similar a.a. sequences fold into similar 3D structures, and (2) the higher the sequence similarity in the conserved protein "core" region, the higher the similarity between their 3D structures *(13)*. Given the sequence of a protein of unknown structure (target), if another protein (template) whose 3D structure has been experimentally determined can be detected, via a.a. sequence similarity to the target, in available DBs, then the 3D structure of the target can be modeled on the basis of the template structure. Since CM is the method that results in the most detailed and accurate protein structure predictions in blind tests *(14)*, it is the elective protein structure prediction method, whenever applicable.

In the next category, FR methods exploit the observation that during evolution protein structures are more conserved than sequences; although proteins with similar sequences have similar 3D structures, similar 3D structures can also be assumed by proteins with relatively different a.a. sequences. Therefore, even if a target protein does not show recognizable sequence similarities with proteins of known 3D structure, its fold might still be similar to one of them. To identify the compatibility of the target sequence with known 3D structures, FR methods take advantage of structural information derived from statistical analyses of protein structure DBs, such as frequency of pairwise a.a. interactions and residue propensity to assume a certain type of secondary structure and/or to be solvent accessible or buried *(15–17)*. Although they have been remarkably successful at identifying suitable structural templates for target sequences without detectable sequence similarities to proteins of known structure, FR methods have two major drawbacks: (1) they are not always able to discriminate between structures truly similar to the target and those unrelated to it, the correct ranking of which remains a challenge; and (2) they are somewhat less successful in recognizing conserved regions between target and template, often producing poor-quality sequence alignments from which to produce the 3D model. As a consequence, these methods should only be used when CM methods are unable to provide an answer.

Both CM and FR methods involve the identification of a suitable structural template in the *PDB* and differ only in the way they detect it (i.e., based on target-template sequence similarity vs. target sequence-template structure compatibility). Once a suitable template has been identified, the procedure used to build the model is essentially the same for both categories. CM and FR are often grouped together in the broader category of template-based protein structure prediction methods and were evaluated in this new category in the most recent *CASP7*.

Conversely, the remaining category of NF prediction methods does not use whole structural template proteins from the *PDB*. However, the most successful of these methods do exploit

information contained in *PDB* structures at a local level. Protein structures contain smaller sub-structures, or structural motifs, which assume the same conformation in the context of different 3D structures. Therefore, it is possible that a "new" fold, i.e., a protein whose 3D structure differs from already known structures at a global level, is formed by a combination of sub-structures that are similar to those present in known structures. Most current NF methods (called fragment-assembly or fragment-based methods) try to reconstruct the global structure of the target by combining structural fragments having a.a. sequence similar to an equivalent short segment of the target sequence, and applying a scoring function to evaluate the resulting models *(18, 19)*. Although at a global level each target sequence generally assumes only one 3D structure, short peptide sequences can assume different conformations depending on their structural context. Therefore, for each target, fragment-based prediction methods have to explore the structure-space that can be occupied by both different fragment conformations and the possible combinations of these fragments. This results in the generation of many models, frequently notably different from one another. As was a previously discussed challenge for FR methods, perhaps the biggest drawback of NF methods lies in their limited ability to discriminate between correct and incorrect models. Nevertheless, these methods represent one of the biggest innovations that have taken place in the field of protein structure prediction for at least a decade, and are the only currently available tool to obtain, however non-systematically, 3D models of small protein structures with new folds. Even when the correct or most-accurate model cannot be identified based on the score provided by the method, having a few potential models can help in experimental structure determination by x-ray crystallography using the Molecular Replacement technique (*see* **Volume I, Chapter 3**), especially in cases of proteins or protein complexes that are proving particularly tricky to solve *(20)*. Additionally, in *CASP* experiments, fragment-based methods have proved to be particularly successful in the prediction of difficult FR targets (*see* **Sections 3.3.1.2** and **3.4**).

The NF category also comprises *ab initio* methods. Contrary to CM, FR, and NF fragment-based methods, *ab initio* algorithms do not exploit information contained in the *PDB* either at a global or local level. Instead, they try to reproduce the physical laws governing protein folding starting only from the a.a. sequence of the target protein and empirical energy functions based on physicochemical principles. Although addressing an intellectually challenging and ever-stimulating problem, until now *ab initio* methods have not produced protein structure predictions competitive with those provided by the methods discussed in the preceding. The practical applications of *ab initio* methods in protein structure prediction are currently limited to short protein segments

(e.g., loops), which cannot be predicted by other methods, and energy refinement of 3D models or parts of them.

One of the remarkable features of protein structure prediction, which has contributed greatly to its diffusion and progress, is that the whole process can be performed using tools that have been made freely available by their foresighted developers. The large majority of the programs and servers described in this chapter can be freely downloaded from, or used through, the Internet (*see* **Table 2.1**).

2. Systems, Software, and Databases

The variety of programs and databases used in protein structure prediction is large and ever-increasing, and a comprehensive listing goes beyond the scope of this chapter. For a full and up-date listing, the reader can refer to:

- The special database and Web server issues of *Nucleic Acids Research*, each published once a year *(21, 22)*, which are dedicated to the most established and popular, as well as recently developed, databases for the collection of biological data and software for their analysis, respectively.
- The special issue of *PROTEINS: Structure, Function, and Bioinformatics (1–6)*, published every 2 years and dedicated to the results of the previous year's *CASP* experiment, comprising both articles describing the evaluation of the performance of protein structure prediction methods in blind tests and articles about the most successful and/or innovative prediction methods.
- The *Livebench* and *EVA* Web sites, in which the performance of automated servers for protein structure prediction on newly released *PDB* targets is assessed on a continuous basis.

Many different types of sequence DBs are available from the *NCBI*, *EBI*, *Sanger Centre*, and *TIGR* Web sites. The most commonly used sequence DBs are central repositories in which sequences from many different DBs are collected, such as the comprehensive non-redundant (NR) protein sequence database at the *NCBI (23)* or *UniProt*, containing accurate annotations about protein function, subcellular localization, and/or other protein features *(24)*. For particular purposes, specialized DBs can be searched, dedicated to specific protein groups or families (e.g., carbohydrate-active enzymes (*CAZy*) *(25)*), or individual genomes (e.g., *Homo sapiens (26)* and *Schistosoma mansoni (27)*), some of which might not be sufficiently complete or refined to be included in the central repositories.

The central repository for macromolecular structures is the *PDB*. DBs derived from the *PDB* and containing classifications of proteins according to their 3D structures and, in some cases, evolutionary relationships (e.g., *SCOP (28)*, *CATH (29)*, *FSSP (30, 31)*), and structural alignment DBs (e.g., *Homstrad (32)*) are also of central importance in protein structure prediction.

The majority of the programs required for protein structure prediction can be run on remote servers accessible through the World Wide Web; therefore, any system supporting a Web browser may be used. Many of these programs are also available to be downloaded and run on most common operating systems (i.e., Linux, Unix, Mac OS X, and Windows). Using programs through the Internet is the easiest option, and the best choice for most common applications. On the other hand, downloading the programs permits a greater flexibility in the choice of parameters and amount of computer power to devote to their use. In turn, this allows for greater automation and removes any reliance on the server availability (a number of protein structure prediction servers are not accessible during the "*CASP* prediction season", which runs from June to September every even year).

The Methods section describes protein structure prediction procedures that take advantage of frequently used methods and programs, highlighting which of them have been performing best in blind protein structure prediction experiments.

3. Methods

Protein structure prediction methods are described in the order: CM, FR, and NF, in agreement with the accuracy of the predictions that they provide in blind tests (from highest to lowest).

3.1. Obtain the Protein Sequence

The starting point of any prediction procedure is the a.a. sequence of the target protein, preferably in *Fasta format*, which is accepted or required by most programs and servers. Servers that do not accept *Fasta format* usually require just a sequence of a.a. residues in one-letter code (equivalent to *Fasta format* without the first header line).

The sequence can be obtained from the *NCBI* Web site or other sequence DBs using similar procedures. (For a simple and comprehensive explanation on the use of the *NCBI* data retrieval system, see the *Entrez tutorial.*) Other programs, such as those available from *Expasy (33)*, *BCM Search Launcher (34)*, or *Sequence Manipulation Suite (35)* perform useful operations on biological sequences, such as the conversion of different sequence formats to *Fasta format* (*ReadSeq*) and of a.a. sequences from

three- to one-letter code (*Three to One*), or the translation of nucleotide gene sequences to the a.a. sequence of their protein products (*Translate*).

3.2. Prediction of Domain Architecture, Transmembrane and Disordered Regions, and Secondary Structures

Most proteins, especially large ones, are comprised of two or more structural units called domains, joined to each other by linker peptides. In case of multi-domain proteins, homologous template domains of known structure may only be available for some of the target domains, they may belong to different proteins whose domain architecture (i.e., the type and linear order of domains comprised in a protein sequence) is different from the target, and/or have a different degree of evolutionary distance from the target domains, so that different techniques may be required to predict their structures. For these reasons, protein structure prediction is generally performed (and assessed, in experiments such as *CASP*) at a domain level. Predicting the spatial arrangement of interacting domains relative to one another in multi-domain proteins (docking) is still a difficult problem, which is only possible to solve in particular cases (see, for example, the results of the Critical Assessment of PRediction of Interactions (*CAPRI*) experiment) *(36, 37)*, and are not addressed in this chapter.

To perform structure prediction at a domain level it is necessary to obtain an initial picture of the domain composition of the target protein and of regions typically found outside globular domains, such as signal, linker, low-complexity, disordered, and transmembrane regions, as well as of the SSE contained in each domain.

1. Domain boundaries are predicted using a variety of approaches based on multiple sequence alignments (MSAs), frequency of amino acids found in domain and linker regions, domain size, predicted secondary structure, sequence comparison methods, neural networks, hidden Markov models, and, sometimes, even using FR and NF methods to build 3D models from which domain definitions are derived (see *(38)* and references therein). *Rosettadom* and *Ginzu (39)* were the most successful automated servers at predicting domain boundaries in *CASP6*, whereas *Phyre* and *DomCut (40)* were the programs used by the best human predictors *(38)*. Other successful groups used *PPRODO (40)* and *CDD (41)*. In *CASP6* the best methods were able to correctly assign domains to >80% of the residues of multi-domain targets. The prediction quality was shown to decrease with an increase of the number of domains in the target, from targets assigned to the CM category to those assigned to FR and NF and for domains made by segments that are not contiguous in the sequence. However, since the number of targets available for domain prediction was relatively small (63 in total, of which about half contained one domain only), the

trends observed in this experiment might not be representative of more general scenarios *(38)*. If the target domains are homologous to those already classified in specialized DBs, servers such as *Pfam (42)*, *SMART (43)*, *ProDom (44)*, *InterPro (45)*, *Prosite (46)*, and *SUPERFAMILY (47, 48)* can provide reliable domain assignments, as well as predictions about signal peptides, low complexity, and transmembrane regions.

2. Natively disordered proteins or protein regions are those that do not assume a regular secondary or tertiary structure in absence of binding partners. Proteins containing relatively long (>30 a.a.s) disordered segments are relatively common, especially in higher eukaryotes, and often have important functional roles *(49)*. The available methods to predict disordered protein regions are based on experimental data from x-ray crystallography and NMR spectroscopy and in general they use machine learning approaches such as neural networks and support vector machines *(49)*. In *CASP6* the best performing method was *VSL2 (50, 51)*, which correctly predicted 75% of disordered residues (true-positives) with an error rate of 17% (false-positives, i.e., ordered residues incorrectly predicted as disordered). Among the other best performing methods were *DISpro (52)*, *IUPred (53)*, *SPRITZ (54)*, *DISOPRED (55)*, and *DRIPPRED*, with about 50% of correctly predicted residues for an error rate <20% *(49)*. A drawback of these methods is that the probabilities associated with the disorder predictions are not always good indicators of the prediction accuracy (48). Assessment of disorder predictions in *CASP6* was confined to targets whose structures were determined mostly by x-ray crystallography, which may impose some order to regions that would be disordered in solution, and whose disordered segments were often rather short *(49)*. Therefore, although the results of this assessment are indicative of the performance of the methods on proteins similar to the *CASP6* targets, they do not necessarily reflect their performance on different types of disorder, e.g., their ability to identify entirely disordered proteins or disordered regions as measured by other experimental methods such as NMR spectroscopy *(49)*.
3. Transmembrane (TM) regions are predicted by *Psi-Pred (57)*, which also provides secondary structure predictions (see the following); *TMHMM (58)* and a number of servers accessible through the *PredictProtein* metaserver *(59)* specialize in predicting transmembrane alpha-helices, and *PRED-TMBB (60)* and *TBBpred (61)* transmembrane beta-barrels.
4. A number of programs have been developed to predict the SSE present in a target sequence at a residue level, and the

best of these have now reached a remarkable level of accuracy. In the continuous benchmarking system *EVA*, several methods are tested for their ability to correctly predict the secondary structure of target proteins showing no significant sequence identity to any protein of known structure. Currently tested methods have been subjected to the evaluation for several months on sets of tens to hundreds of proteins, and were shown to have an accuracy, defined as their ability to correctly predict each of the target residues as being in alpha-helical, beta-strand, or other (coil) conformation, of at least 70% on their overall set of target proteins. This accuracy depends on the quality of the MSA (*see* **Section 3.3.2.1**) that is automatically built for the target, ranging from about 64% for a single sequence with no homologs to nearly 90% for targets with high-quality MSAs. In addition to the high prediction accuracy they achieve, another important feature of these methods is that the confidence values associated with their residue-based predictions correlate well with the actual accuracy of the prediction. Several consistently well-performing methods, such as *PROF (59)*, *Psi-Pred (57, 62)* and *SAM-T99 (63)* predict correctly >76% of residues of the set of targets they have been tested on. *Porter (64)* achieves almost 80% correctly predicted residues and is currently the best among the methods evaluated by *EVA*, but as its evaluation started more recently than that of the other methods it has been tested on a smaller set of target proteins. Another useful server is *Jpred (65, 66)*, which provides consensus SSE predictions between several methods.

In general, searches with any of these servers are easy to run and do not take long to complete (from a few minutes to a few hours, depending on the server); therefore, it is advisable to run the target sequence through more than one server to get as accurate as possible a starting guess about the target domain boundaries and SSE composition. Ideally, since the confidence in a given prediction is higher when different methods are in agreement, it is advisable to use several methods and compare their predictions, especially when the methods performance in blind tests has not yet been assessed, the assessment has been performed on a small number of targets, or the target has significantly different features from those on which the assessment has been performed (e.g., it contains large disordered regions).

All the predictions from now on should be performed using as input single domain sequence regions as opposed to whole protein sequences. Given the uncertainty in domain boundary prediction, it is advisable to include in the target domain sequence some 10–20 a.a.s N-terminal and C-terminal to the boundaries predicted by domain prediction servers. A more precise prediction of the domain boundaries can often be obtained at a later

stage, following the production of optimized target-template sequence alignments and 3D target models.

3.3. Template-Based Modeling

Template-based modeling consists of the following key steps: (1) identification of the template structure; (2) refinement of target-template(s) sequence alignments; (3) model building; (4) model evaluation; and (5) model refinement.

3.3.1. Identification of the Template Structure

3.3.1.1. Sequence Comparison Methods

Sequence comparison (SC) methods are used to retrieve proteins similar to the target from sequence DBs. Proteins that score better than certain threshold values are considered to have statistically significant sequence similarity to the target, based on which an evolutionary relationship (i.e., homology) with the target is inferred. Since similar sequences fold into similar 3D structures, proteins of known structure scoring above the threshold can be used as template(s) from which a comparative model of the target is built. Homology models are usually built using as main template the structure of the protein having the highest global sequence conservation with the target. However, regions containing insertions and deletions in the sequence alignment of the target with the best global template might be better conserved, at a local level, in other homologous proteins of known structure, which can therefore be used as templates for these regions. In principle, homologous proteins may be retrieved searching sequence DBs including only proteins of known structure (e.g., pdb at the *NCBI*). However, since protein homologs of unknown structure provide important information for model building, and the sequences of proteins of known structure are also contained in comprehensive sequence DBs such as *NR*, these are generally the DBs of choice.

SC methods can be assigned to two main categories: pairwise methods and profile-based methods. Pairwise sequence comparison methods (e.g., *BLAST*) *(67)* compare the sequence of the target with each sequence present in a sequence DB. They are able to detect proteins showing high sequence similarity to the target, based on which the target and the retrieved proteins are inferred to be close evolutionary relatives and expected to assume very similar 3D structures. Profile-based sequence comparison methods (e.g., *PSI-BLAST*) *(67)* compare each sequence in a DB with a profile created from an MSA of the target and its closest homologs, which have been previously detected using pairwise methods. Since the profile incorporates information about several family members and not just the target, these methods are able to detect proteins evolutionarily related to the target that cannot be detected by pairwise sequence comparison methods. Based on their lower sequence similarity with the target, these proteins are considered to be distant evolutionary relatives; therefore, their structural similarity with the target might be lower than that

of proteins matched by pairwise methods (however, this is not always the case, *see* **Section 3.3.1.2** and **Fig. 2.2**). In *CASP6*, if a significant match with a correct template domain (as defined based on the structural similarity of the target with structures in the *PDB* measured by the *LGA* program, *see* **Section 3.3.2.4**) was found using *BLAST*, the target was assigned to the "easy" CM prediction sub-category; any targets for which structural homologs were detected by *PSI-BLAST* but not by *BLAST* were considered to be "hard" CM targets *(14, 68)*.

The most popular pairwise sequence comparison methods are *BLAST*, *FastA (69)*, and *SSEARCH (70)*. *BLAST* is the most commonly used, it is implemented by almost every sequence DB available on the Internet and can either be used interactively or downloaded, together with a number of comprehensive or specific sequence DBs, from the *NCBI* Web site. Although installing *BLAST* on a local computer can be advantageous to run multiple searches automatically and allow for greater flexibility in parameter settings, the Web versions are easier to use and provide reasonable default parameters, which might be preferable for first-time users. *BLAST* returns pairwise alignments of the target with sequences retrieved from the DB, and several parameters to help decide about the significance of each alignment, i.e., whether the matched protein is likely to a be a real homolog of the target as opposed to showing sequence similarity with it purely by chance.

1. The expectation value (E-value) of an alignment represents the number of different alignments with scores equivalent to or better than the score of that alignment (**Note 2**) that are expected to occur by chance in a database search. Therefore, the lower the E-value, the higher the probability that a matched sequence is a real homologue of the target, and vice versa. Unfortunately, there is no universal threshold value that guarantees identification of all the true homologs and rejection of all non-homologous sequences. In general, for a given threshold value, there will be both proteins with E-values better (i.e., lower) than the threshold that are not real homologs of the target but show some sequence similarity with it purely by chance (false-positives), and proteins with E-values worse (i.e., higher) than the threshold that are homologous to the target but have diverged from it to the point that the sequence similarity is not distinguishable from that occurring by chance (false-negatives). Lowering the E-value threshold results in a decrease in the number of false positives (i.e., incorrect hits) and an increase in the number of false-negatives (in that a higher number of real homologs have E-values above the threshold, and are discarded). Conversely, increasing the E-value threshold results in a lower number of false-negatives (i.e., missed hits) and a

higher number of false-positives (e.g., for E-values of 10 or higher, a considerable number of hits found by chance have E-values below the threshold, and are selected). In general, a match is considered to be significant if the E-value is around 10^{-2}–10^{-3} or lower, whereas E-values of 10^{2}–10^{3} or higher indicate that the match is almost certainly not to be trusted. Matches with E-values between or approaching these values should be evaluated carefully taking into account other parameters.

2. The percentage of sequence identity (%_ID) represents the number of identical a.a.s found at corresponding positions in the aligned regions and it can provide indications about the homology between two proteins. For sequences aligning over about 100 a.a.s, a %_ID above 40% indicates certain homology, whereas a %_ID of 20% or less might occur purely by chance; if the %_ID is between 20 and 40% the two proteins might be homologous, but additional information is required to support this hypothesis. These threshold values vary with the length of the aligned regions. Since short segments have a higher chance of showing sequence similarity by chance, for considerably shorter and longer alignments the %_ID required to infer homology is therefore higher and lower, respectively. Similar information to the %_ID is provided by the percentage of sequence similarity, which depends on the substitution matrix used by SC methods (**Note 3**). High values of percentage of sequence similarity qualitatively support the significance of the match and the correctness of the alignment, but the relationship of this parameter with homology is even less precise than for %_ID.
3. Between closely related proteins, insertions and deletions (gaps) are usually relatively few and generally cluster in surface loop regions, rather than being spread all over the structure and interrupting regular SSE. Therefore, the lower the number of gaps and different positions in which they occur in the alignment, the higher the significance of the match and quality of the alignment.

In the absence of overall sequence similarity, i.e., in case of E-values higher and %_ID lower than the aforementioned values, the following additional sources of evidence can support the hypothesis of the existence of evolutionary relationships between sequences retrieved from the DBs and the target: the similarity of both the target and the retrieved sequence to a third "intermediate" sequence that is more closely related to each of them than they are to each other; the existence of structural relationships between the matched proteins of known structure (as shown, for example, by their classification within the same *SCOP* Family, Superfamily or Fold) (**Note 4**); a good overlap between the SSE of the templates and the SSE predicted for the target

(*see* **Section 3.3.2.2**); and finally the conservation of residues that are known to play important structural and/or functional roles for the target and/or the retrieved proteins (key residues, *see* **Section 3.3.2.5**). Additionally, since DB searches are not symmetrical, confidence in the target homology with a retrieved sequence can be increased if the target sequence is matched, in turn, when searching the DB using the previously retrieved sequence as a query.

The availability of biological information about the target family and experience with alignments from many different protein families can also be of help in the evaluation of the biological significance of a match.

If proteins of known structure closely related to the target are not found by pairwise methods, it is possible to change the aforementioned parameters to tailor them to the specific problem at hand. As an example, to detect distantly related proteins it is possible to choose a lower number BLOSUM or higher number PAM matrix (*see* **Note 3**), and/or decrease the penalty associated with the insertion and elongation of gaps in the alignment (*see* **Note 2**). However, a more efficient way to detect distant homologs *(71)* consists in the use of profile-based methods such as *PSI-BLAST (67)* and HMM-based methods *(63, 72, 73)*.

In *PSI-BLAST (67)* the first iteration, coinciding with a simple *BLAST* search, is used to collect sequences similar to the target based on a pre-defined E-value threshold (a default value is provided, but it can be changed by the user). The alignments of these sequences to the target are used to build a multiple alignment from which a Position-Specific Score Matrix (PSSM), or profile, is derived that contains values related to the frequency with which each a.a. occurs at each alignment position. In the second *PSI-BLAST* iteration, the sequences in the DB are matched to this profile, rather than to the target sequence. If a third iteration is run, the profile is updated to incorporate in the alignment the new sequences found with E-values below the threshold; if no new sequences are found, the profile does not change, i.e., the program has reached convergence. *PSI-BLAST* iterations can be run until the program converges or a protein of known structure is matched below the threshold. *PSI-BLAST* results can be evaluated using the same parameters described for *BLAST*. However, from the second iteration onward *PSI-BLAST* E-values are not directly comparable to those calculated by *BLAST*. Indeed, the E-value associated with a protein retrieved by *BLAST* is different from the E-value associated with the same protein, retrieved from the same DB, by any *PSI-BLAST* iteration following the first. The reason for this is that *BLAST* scores the target sequence against each DB sequence using a matrix (e.g., BLOSUM62) containing fixed values for each a.a. pair, independent of the position where they occur in the sequence alignment, whereas *PSI-BLAST* scores the target sequence against a PSSM whose values

(depending on the frequency of a.a.s observed at each position in the MSA from which the PSSM was generated) are updated after each iteration. Because it is derived from an alignment of multiple sequence homologs, the PSSM is more powerful than the fixed scoring matrices, and can give sequences homologous to the target a higher score and therefore a better E-value, thus promoting them over incorrect matches. However, while convergence is rarely reached, if sequences non-homologous to the target are matched with E-values below the defined threshold (false-positives), they will be incorporated in the *PSI-BLAST* profile leading to the matching of more non-homologous sequences in the following iteration, and ultimately to divergence from the original target. Indeed, the profile can drift away from the originating target sequence so far that eventually the target sequence itself, and not only its homologs, will score very poorly against the profile! To prevent this, the hits collected by each *PSI-BLAST* iteration should be carefully examined, adjusting the threshold for inclusion (to make it more restrictive in case of divergence and more permissive in case convergence is reached before a protein of known structure is matched) and/or selecting manually sequences to be included in or excluded from the profile.

If *PSI-BLAST* does not identify any sufficiently convincing hits, or converges before identifying any matches to proteins of known structure, hidden Markov model (HMM)-based programs can be used *(74)*. HMMs can also be run to detect additional templates and/or compare their results with those obtained by *PSI-BLAST*. Starting from an MSA, these programs build a hidden Markov model (HMM) that, similarly to a *PSI-BLAST* profile, represents the properties of all the sequences in the alignment. This HMM is then used to search the DBs for homologous proteins. The two most popular HMM-based programs are *SAM* and *HMMER (72, 73)*, both freely available for downloading. *SAM* is also accessible through a Web server interface (*SAM-T02*) *(75, 76)* that takes the target sequence as input and automatically builds both the MSA and the HMM. Although expert users might prefer to use the downloadable version to be able to modify program parameters, the Web interface is straightforward to use and provides results relatively quickly. The *SAM-T02* output provides E-values (i.e., estimates of approximately how many sequences would score equally well by chance in the database searched) and the *SCOP* classification of the matched structures to help evaluate the matches. If the E-values are higher than the suggested significance threshold (e.g., E-values $<10^{-5}$ and higher than 0.1 indicate very reliable and speculative matches, respectively) and/or proteins matched by the HMM do not belong to the same *SCOP* superfamily (**Note 4**), additional information is required to infer homology between any of the matched proteins and the target (see the preceding). To speed-up the search for homologs,

several methods have been developed that allow comparison of the target sequence with pre-calculated profile libraries, such as PSSMs generated with *PSI-BLAST* (*IMPALA*) *(77)* and HMM libraries representing homologous sequences (*Pfam* and *SMART*) or proteins of known structure that are evolutionarily related at a superfamily level as defined by *SCOP* (*SUPERFAMILY*).

More recently, a number of profile–profile comparison methods have been developed capable of detecting distant relationships that are not recognized by sequence-profile matching methods *(78)*. These include: prof_sim *(79)*, *Arby (80, 81)*, *COMPASS (82–84)*, *Picasso (85)*, *COACH (86)*, *LAMA (87)*, *PRC* (the profile–profile implementation of *SUPERFAMILY*), *FFAS03 (88)*, *FORTE (89)*, *ORFeus (90)*, and *HHpred (91)*.

The most successful prediction groups in the last *CASP* editions used *SAM-T02*, *FORTE*, *ORFeus*, and *FFAS03*, either as stand-alone programs or as part of metaservers (***see*** **Section 3.5.2**). The performance of all these methods, together with that of *HHpred*, *SUPERFAMILY*, and *PRC* is also subjected to continuous evaluation by the *Livebench* server.

Since several sequence–profile and profile–profile comparison methods (e.g., *SAM-T02*, *ORFeus*, and *HHpred*) exploit structural information (most often, secondary structure predictions and secondary structure assignments by *DSSP (92)*), sometimes they are classified together with FR methods or metaservers. Targets for which a correct template structure (i.e., a structure similar to the target according to the *LGA* program, *see* **Section 3.3.2.4**) was identified in the *PDB* by profile–profile sequence comparison methods were assigned to the FR/H (H: homologous) category in *CASP6*; conversely, targets for which the correct template structure could not be detected by any sequence-based methods was assigned to the FR/A (A: analogous) category *(68)*.

3.3.1.2. Fold Recognition Methods

As mentioned in the Introduction, analysis of protein sequence and structure DBs led to the observation that, as proteins diverge, overall structural similarity persists even when no significant sequence similarity can be detected. In fact, two proteins with <25–30% overall identities can have either very similar or completely different 3D structures. In order to detect an evolutionary relationship in the absence of sequence similarity, FR methods: (1) identify potential structural similarity signals within the sequence, and (2) apply confidence statistics to rank potential matches and provide confidence values for the prediction in order to distinguish "real" matches (true-positives) from spurious unrelated ones (false-positives). FR methods try to assess the likelihood of the target proteins sharing a fold with one of the proteins of known structure by comparing structural features predicted for target sequences, on the basis of statistical analysis of known

protein structures, with those actually observed in each structure. Structural features commonly taken into account include 3D environmental preferences of each amino acid, such as propensity to form pairwise contacts with other residues, solvent accessible surface area, and local secondary structure. This comparison between one-dimensional (1D) sequences and 3D structures is usually done by either encoding 3D information into 1D sequences *(15)* or by threading, i.e., inscribing the 1D sequence of the target into each fold contained in a library of representative 3D structures *(93)*. The compatibility of the target sequence with each fold is evaluated using empirical or "knowledge-based" potentials that take into account the aforementioned structural properties.

The Three-Dimensional Position Specific Scoring Matrix (*3D-PSSM*) *(94)* server uses sequence and structure alignments, as well as secondary structure and solvent accessibility, to construct descriptive position-specific matrices for each domain in a non-redundant structural DB. These matrices can be compared with PSSMs or profiles of a query sequence and the results reported with underlying evidence to enable the user to understand the strength and confidence of the fold prediction. *3D-PSSM* and its eventual successor *Phyre*, both have simple and user-friendly Web-based interfaces. The calculations can take some time as the many-by-many profile comparisons can be intensive. Eventually the results will be displayed on a Web page which is e-mailed to the user. The page displays proteins of known structures that are predicted to fold in a similar way to the query sequence. These proteins are listed in a table in order of predicted similarity to the query, so that the proteins at the top of the table are predicted to be most similar. An E-value is provided to indicate the probability that the prediction is true, the lower the E-value the more likely the prediction is to be correct. The alignment to the target sequence and *SCOP* classification of the protein structures are provided. *3D-PSSM* also provides an automatically generated homology model based on the alignment to each of the matched structures. If the steps for using the server are simple, most attention must be paid to interpreting the results returned. With all prediction methods, a sign of a prediction being correct is whether it is persistent. For example, the confidence in the prediction can be increased if the top reported *3D-PSSM* structures matching the query belong to the same *SCOP* family or superfamily, or even fold (**Note 4**), and/or if running through the server sequences homologous to the target the same structures appear on top of the list. Conversely, if the top hits belong to different families or superfamilies, a fold similar to the target might still be present among them, even if the method cannot identify it clearly from the wrong ones. In such cases, to detect the correct fold from incorrect ones, and to

further support FR predictions in general, it is possible to exploit additional information, such as those used to validate the output of SC methods (*see* **Section 3.3.1.1**). As an example, residues strongly conserved in the MSA of the target sequence can be mapped on the known structures to check whether they play key structural or functional roles (e.g., involvement in disulfide bridges, building of active sites, etc.).

There are a number of fold recognition methods available via the internet. *3D-PSSM*, *Threader (16)*, *GenThreader (95)*, and *mGenThreader (96)*, *TASSER-Lite (97)* (combining the PROSPECTOR 3.0 threading algorithm *(98)* with fragment-based NF prediction methods), *Fugue (99)*, *RAPTOR (100)*, *HMAP (101)* (using structure-based profiles in the conserved regions and sequence-based profiles in the loops), *LOOPP* (102), *SPARKS2 (103, 104)*, and *SP3 (104, 105)* have been used as standalone programs or as input for metaservers (*see* **Section 3.5.2**) by the best performing group in the CM and FR categories in the last *CASP* editions, and their performance, together with that of many other servers, is continuously assessed by *Livebench*. In general, metaservers that use a consensus of SC and FR-based methods are most successful at predicting FR targets that are clearly homologous to proteins of known structure, classified in the FR/H (H: homologous) sub-category in *CASP6*. Conversely, the best predictions of FR targets assigned to the *CASP6* FR/A (A: analogous) sub-category, for which no clear evolutionary relationship with already known folds can be detected, are provided by fragment-based methods (*see* **Section 3.4**).

3.3.2. Refinement of Target-Template(s) Sequence Alignments

Together with the extent of structural similarity between target and template, the generation of structurally correct target-template sequence alignments (i.e., sequence alignments corresponding to the optimal structural superposition between the target and template structures) is one of the most important factors affecting the final quality of template-based models. Therefore, sequence alignment(s) provided by SC and/or FR methods should be critically evaluated and refined, with increasing care as the %_ID between target and template decreases. Because of the difficulty of obtaining a correct alignment, many of the best performing groups in *CASP6* generate and evaluate both a number of target-template sequence alignments obtained from different sources and/or using different alignment parameters, and a number of 3D models produced by different servers. The evaluation is based on the results of model quality assessment programs (MQAPs) (*see* **Section 3.3.4**) and/or on the assumption that consensus predictions provided by independent methods are more reliable than any single prediction; accordingly, consensus regions are taken as such while variable regions are re-aligned and subsequently re-evaluated in a cyclic procedure.

3.3.2.1. Multiple Sequence Alignments

One way to refine pairwise target-template alignments is by comparing them to the alignments produced by MSA methods (*see* **Volume I, Chapter 7**). MSA programs align all sequences in a given set to one another, producing MSAs that can be more accurate than those produced by SC or FR methods. *ClustalW (106)* and *T-Coffee (107)* are among the most widely used programs to build MSAs. Both are available for downloading (a windows interface version of *ClustalW* called *ClustalX (108)* is also available), and can be used interactively through a Web interface. However, several newer and potentially more powerful methods are now available (*see* **Volume I, Chapter 7**). Together with the specific features of the MSA program used, the set of sequences given as input to the program is one of the most important factors affecting the quality of the final MSA.

MSAs for protein structure prediction are typically generated from sequences putatively homologous to the target identified by SC methods. Sequences matching the target with E-values significantly higher than that of the selected threshold or of the most distantly related of the selected templates might be eliminated in that they should not contribute to improving the quality of the target-template(s) alignments, and might actually make it worse. However, since false-negatives might be among them, these sequences can also be kept and their relationships with the target evaluated at a later stage, on the basis of the resulting MSA itself. The pattern of conserved residues in an MSA can provide information about key structural or functional residues in a protein family and increase the confidence in the existence of evolutionary relationships between target and template (*see* **Section 3.3.2.5**). Potential templates detected by FR methods may be added to the set of target homologs retrieved by SC methods and given as input to MSA programs; however, they often have sequences too different from the target and its homologs to produce good alignments. In such cases, the sequences of the putative FR templates can be used as queries by SC methods to retrieve their homologs from sequence DBs and MSAs can be produced for the templates as well. Comparison of the MSAs produced for the target and template sequences, and in particular of the pattern of conserved and putatively essential residues, may support the hypothesis of the existence of evolutionary relationships between them. An informative way to highlight sequence conservation in MSAs is provided by the *WebLogo (109)* program, producing a picture in which residues occurring at each alignment position are shown by the one-letter code, the size of each letter being proportional to their frequency of occurrence. MSAs are most informative when they contain a relatively high number of sequences, similar enough to one another to be certain of their homology and to allow for the production of a correct alignment, and divergent enough to allow for conserved

positions to be distinguished from variable ones. For this reason, redundant and outlier sequences (e.g., those that have %_ID >80% with any other sequence in the alignment or <20% with all other sequences in the alignment, respectively) are usually eliminated. Other kinds of editing include the deletion of alignment regions other than those matching the target domain, which may be present, for example, if the sequences retrieved from the DBs comprise additional domains besides that homologous to the target. Additionally, shifts in the aligned sequences can be introduced manually based on structural information such as those described in the following (**Sections 3.3.2.2, 3.3.2.4,** and **3.3.2.5**). Although guidelines such as these can be useful to start with, production of a good-quality MSA is very much a process of trial and error. In general, choices on how to edit MSAs depend on specific features of the sequences contained therein, and on a balance between the computational and human time required to analyze a high number of sequences and the accuracy requested of the final result. This, in turn, depends largely on the difficulty of the prediction: in case of "easy" targets, for which reliable templates aligning well with the target can be identified, even the pairwise *BLAST* alignments might be sufficient; conversely, when the templates are distant homologs, detectable only by profile-based or FR methods, with very low %_ID and difficult to align, all available information from sequence and structural homologs should be exploited. Often, several rounds of alignment editing and re-alignment are required to produce high-quality MSAs.

Several programs are available to visualize MSAs, allowing the user to color a.a.s according to residue type or conservation, edit the alignment to eliminate redundancies and outliers as well as alignment columns and blocks, and save them in different sequence formats. Such programs include *Belvu (110)*, *Seaview (111)*, *BioEdit*, *CINEMA (112)*, and various other tools from the *Expasy* Web site.

3.3.2.2. Comparison Between SSE Identified in the Templates and Predicted for the Target

SSE are usually among the best conserved parts of evolutionarily related proteins; therefore, they should also be found in corresponding positions in the sequence alignment. Additionally, insertions and deletions of more than one or two residues are unlikely to occur within SSE, whereas they can easily take place within the conformationally more variable and solvent exposed loop regions.

The secondary structure assignment for the template structures can be calculated by programs such as *DSSP* or obtained from the *PDB* Web site. This contains both the *DSSP* automated assignment and the manual assignment provided by the experimentalists who have solved the structure, which may be more accurate than those calculated automatically. Secondary structure predictions for the target can be obtained as described above (**Section 3.2**). Mapping the SSE calculated for the template(s) and predicted for the target on the target-template(s) pairwise alignments or MSAs might help

refine the sequence alignments. If the SSE do not align and/or large insertions or deletions occur within these regions, the target-template alignment may be modified to adjust for these features, which are likely to be incorrect. However, unless it is obvious how to correct errors in the sequence alignment, modifying it by hand is only advised once one has reached a certain degree of experience. Hybrid- or meta-profiles combining sequence and secondary structure information are now used by many of the methods that have been most successful in the *CASP* experiments (e.g., *Meta-BASIC (113)*, *see* **Section 3.5.2**); additionally, several FR methods (e.g., *3D-PSSM*, *Phyre*, and *mGenThreader*) report the SSE of the templates and those predicted for the target in their output target-template alignments.

3.3.2.3. Identification of Domain Boundaries in the Template Structures

The template region aligned to the target by SC methods does not necessarily correspond to a structural domain of the template, but it can be shorter, if only a fraction of the template domain displays recognizable sequence similarity with the target, or longer, in case regions before or after the template domain have also similar sequence to the target. Indeed, the initial domain definition of the target, which has been preliminarily provided by domain prediction servers (**Section 3.2**), may be refined during the prediction procedure based on accurate target-template(s) alignments. Until this step is performed, it is advisable to use a slightly larger segment of the target sequence than that predicted to correspond to the target domain.

The boundaries of the template domains matching the target can be identified based on *SCOP*, which is widely believed to be the "gold standard" of protein structure classification, and where protein domains are classified according to sequence, structural, functional, and evolutionary criteria. In case the template structures have been made available from the *PDB* more recently than the date of the latest *SCOP* release, a domain assignment for them may be found in other structural classification DBs, such as *CATH* or *FSSP*. If no DB contains a pre-calculated domain definition for the selected templates, this can be obtained by running the template structures through one of the available programs for domain assignment, such as *Domain Parser (114, 115)*, *PDP (116)*, or *DOMPLOT (117)*. Mapping the structural domain definition of the templates on the target-template sequence alignment(s) can help to refine the initial prediction of target domain boundaries.

3.3.2.4. Structural Alignment of Selected Template Domains

The structural alignment of the template domains shows which regions are structurally conserved among them, and are therefore likely to be conserved in a target protein of unknown structure evolutionarily related to them. This alignment might be extended, by including other proteins of known structure whose evolutionary relationships with the template(s) have been ascertained on

the basis of structural criteria, to get a more precise definition of the conserved "core" and variable regions among proteins homologous to the target. Proteins of known structure evolutionarily related to the template(s) can be obtained from *SCOP*, in which closely and distantly related homologs are classified within the same family or superfamily, respectively, (*see* **Note 4**), or from other structural classification DBs mentioned in **Section 2**.

Pre-compiled alignments of protein domain structures are available from specific DBs (e.g., *Homstrad*, *FSSP*, and *CE*) and may include the selected templates. Alternatively, structural alignments of the templates can be built by using one of the many available programs for protein structural alignment (a large, if incomplete, list is available from *Wikipedia*). Choosing which one(s) to use depends partially on how many proteins match the template and have to be aligned. Some programs perform only pairwise alignments (e.g., *LGA (118)*, *DaliLite (119)*, *Mammoth (120)* and *VAST (121)*) whereas others can perform multiple structure alignments (e.g., *CE (122)*, *SSAP (123)*, *SSM (124)*, *PrISM (125)*, *Mammoth-mult (126)*, and *POSA (127)*), providing a global alignment of all the input structures and, in principle, better results. Although no automated program is capable of systematically producing accurate alignments of highly divergent structures, most structural alignment programs can produce relatively good alignments of similar structures, as should be the case for templates identified based on sequence similarity to the same target. Depending on the time available and the level of accuracy required, the templates might be run through several servers to compare the results.

Many predictors successful at *CASP* exploit information deriving from the structural alignment of multiple templates to identify conserved regions, guide sequence alignments, and/or build chimeric models from fragments extracted from different templates, to be compared and evaluated in subsequent steps using consensus and/or quality assessment criteria (*see* **Section 3.3.4**).

3.3.2.5. Structural Analysis of Selected Template Domains

Regions that are structurally conserved among the templates and their homologs and, therefore, are putatively conserved in the target structure as well, in general correspond to SSE and conserved loops, and should not contain insertions or deletions in the target-template sequence alignments. Therefore, if "gaps" occur within these structurally conserved regions, they will have to be moved to regions where they might be more easily accommodated from a structural point of view.

As protein structures are more conserved than sequences during evolution, structurally conserved regions may have low sequence similarity; nevertheless, they should contain "key" structural features allowing them to assume similar conformations. Such features have been identified in the past for several

protein families and found to consist, most often, of residues belonging to one of the following categories: (1) residues, in general hydrophobic in nature, having low solvent accessibility, interacting with one another within the protein core (*128*); (2) residues conferring special structural properties to the protein region where they occur: Gly, Asn, and Asp, able to assume more frequently than other residues positive φ values; or Pro, whose main-chain nitrogen atom is involved in peptide bonds found more frequently than those of other residues in *cis*, rather than *trans*, conformation and that, lacking the main-chain hydrogen bond donor capability, cannot take part in the formation of structures maintained by regular hydrogen bond patterns, e.g., α-helices or internal β-strands; (3) Cys residues, that can be involved in the formation of covalent disulfide bonds; (4) any other residues that appear to play a specific structural role in a given family (e.g., negatively charged Asp and Glu and polar Asn residues binding calcium ions in the cadherin family *(129, 130)*, and polar or charged residues able to form hydrogen bonds or salt-bridges maintaining the conformation of antibody hypervariable loops *(131–134)*). Comparison of evolutionarily related structures can allow identifying conserved residues having a key structural role, which are likely to be conserved in the target structure as well, and represent useful landmarks to help refine pairwise target-template sequence alignments and MSAs. *Joy (135)* is a program that reports different types of residue-based structural information (e.g., secondary structure, solvent accessibility, positive φ values, *cis*-peptides, involvement in disulfide and hydrogen bonds) on sequence alignments. Additionally, *NACCESS (136)*, *HBPLUS (137)*, and *Procheck (138, 139)* can be used to calculate solvent accessible surface area, involvement in hydrogen bond formation, and main-chain dihedral angles (Ramachandran plots), respectively.

Residues playing a key functional role are also likely to be conserved, in either type or physicochemical properties, between the target and template structures. These may be identified from the literature and/or from template structures in complex with their ligands, in case these are available. Protein–ligand contacts can be calculated by *LIGPLOT (140)* and pre-calculated contacts can be obtained from the *PDBsum* Web site *(141)*.

In principle, residues highly conserved in MSAs might be involved in key structural and/or functional roles. However, in practice, the possibility to predict involvement in important structural or functional roles from sequence conservation is highly dependent on the MSA "quality": if the sequences comprised in the MSA are too closely related, other positions besides the essential ones will be conserved; conversely, if the sequences are distantly related, essential structural or functional roles might be played by non-identical residues sharing specific physicochemical

features that might not be easy to identify from an MSA. Nevertheless, once putative key structural and/or functional residues have been identified as described, the conservation of specific features (e.g., main-chain flexibility, hydrogen bond donor ability, presence of an aromatic residue at a position involved in a cation-π interaction, etc.) at given positions can be searched in an MSA in a targeted way. Programs like *COLORADO3D* (142) and *CONSURF (143)* color each residue in a protein structure according to its conservation in the MSA given as input and, therefore, allow visualization of the structural location of conserved and variable residues.

3.3.3. Model Building

This section describes the generation of a 3D protein model based on target-template(s) sequence alignment(s).

3.3.3.1. Choice of the Best Overall Template(s)

Choosing the best template is usually done based on both structural features of the template(s) and information collected for the production of optimal target-template(s) alignments as described in **Section 3.3.1.1,** such as: E-values, and other statistical scores provided by different methods to evaluate the likelihood of the existence of structural relationships between target and template(s); %_ID; number and distribution of gaps; length of the aligned regions (i.e., coverage of the target sequence); correspondence between SSE of the templates and those predicted for the target; absence of insertions and deletions in the target-template alignment within regions corresponding to those structurally conserved among the templates; and conservation of key structural and functional residues between target and template.

If good alignments with different templates are available, structural considerations can help decide which template structure(s) is/are likely to produce most suitable models for their intended applications. In general, x-ray crystallography provides more precise pictures of protein structures than NMR spectroscopy. However, it should be kept in mind that x-ray structures and, therefore, models based on them, represent static images of proteins, which often assume multiple conformations in solution. In the case of structures determined by x-ray crystallography, the parameters to take into account are the following (*see* also **Volume I, Chapter 3**).

1. Resolution and B-factor. The lower the values of these parameters, the better the quality of the structure. In general, for resolution values <2.0 Å the quality of the structure is very high, for values >3.0 Å it is low; B-factor values <30–35, in the range 40–80, and >80 indicate well-determined, mobile, and unreliable regions, respectively.
2. Completeness. Regions corresponding to those relevant for our model in the target-template sequence alignment should not be missing from the template structure (e.g., N-terminal

regions are often cleaved out; exposed loops and N- and C-terminal regions might remain flexible in the crystal structure and not be determined; older, low-resolution structures may contain only Cα carbon atoms; etc.).

3. Protein conformation. The same protein domain can have different conformations in different *PDB* files, depending on the functional state of the domain (e.g., free vs. ligand bound), crystal packing interactions (i.e., interactions with other copies of the same molecule in the crystal), experimental conditions used (e.g., pH, ionic strength, etc.). All this information is contained in the coordinate files of the protein structures, which can be freely downloaded from the *PDB* and visualized using a number of freely available structure visualization programs such as *Swiss-PDB Viewer (144)*, *RasMol (145)*, *Protein Explorer (146)*, *Cn3D (147)*, *WHAT IF (148)*, *MolMol (149)*, and *VMD (150)*. The choice of the template(s) is a trade-off between all these different factors, and the purpose of the model may also be a useful guide (e.g., when modeling receptor–ligand interactions, templates in the ligand-bound conformation should be chosen, if available).

In experiments like *CASP*, predictors often build chimeric models assembling together fragments taken from different templates and/or build several models, based on different main templates and different target-template alignments, which are evaluated at a later stage (*see* **Section 3.5.2**).

3.3.3.2. Model Building

Template-based modeling involves taking advantage of as much as possible information from proteins of known structure putatively homologous to the target, i.e., from the selected template(s). Once refined target-template(s) sequence alignments have been obtained and one or more principal templates have been chosen, model building itself is a relatively straightforward procedure, which can be carried out interactively using structure manipulation programs such as *Swiss-PDB Viewer*, *InsightII/Biopolymer*, or *WHAT IF*.

1. Modeling of the main-chain atoms of regions conserved in the template structure(s). If a single best-template has been detected, the main-chain atoms of conserved regions in the optimized target-template alignment are imported from this template. Conversely, if different regions of the target appear to be more closely related to different structures (e.g., they contain 'gaps' in the alignment with the best template but not with other templates), the conserved main-chain regions of the templates are optimally superimposed (using the structural alignment programs mentioned in **Section 3.3.2.4**), and the regions to serve as templates for different segments

of the target are joined together and imported in the target model.

2. Modeling of the main-chain atoms of structurally variable regions. In case insertions and/or deletions are present in the sequence alignment of the target with all templates and cannot be modeled based on the coordinates of homologous structures, several techniques can be applied to model the regions containing them (usually loops). For a restricted number of loops, sequence–structure relationships have been described, based on which loop conformation can be predicted quite accurately (e.g., antibody loops *(131–134, 151)*, and β-hairpins *(152, 153)*). However, with the exception of these particular cases, and although encouraging results have been achieved in *CASP6* in the CM category, in which four groups were able to predict loops larger than five residues with RMSD <1.0 Å *(14)* (*see* **Note 1**), no method is currently available to accurately and consistently model regions of more than five residues that cannot be aligned to a template; the larger the loops, the more difficult their prediction. Therefore, these regions may either be left out of the model, especially if they are far from the sites of interest of the protein (e.g., active sites) or, if included, it should be pointed out that their reliability is much lower than that of the regions conserved in, and imported from, homologous templates.

   One common way to model loops is based on structural searches of the *PDB* database for protein regions having: (1) the same length as the loop to model; (2) a similar conformation of the main-chain atoms of the residues before and after the loop; and (3) a similar pattern of "special residues" that can confer special structural properties to the protein region in which they occur (e.g., Gly, Asn, Asp, or Pro, *see* **Section 3.3.2.5**) and are often important determinants of loop conformation. Conversely, "*ab initio*" methods do not use information contained in structural DBs but try to simulate the folding process or explore the conformational space of the loop region, for example, by molecular dynamics or Monte Carlo methods, followed by energy minimization and selection of low-energy conformations *(154)*.

   The interactive graphics software *Swiss-PDB Viewer* provides options to evaluate the compatibility of loops derived from structural searches of the *PDB* with the rest of the target model based on the number of unfavorable van der Waals contacts that they establish and on the results of energy calculations. The groups using *Swiss-PDB Viewer* and the *Loopy* program *(155)* of the *Jackal* package were among the most successful predictors of loops in the *CASP6* CM category. Other software for loop modeling includes *LOBO (156)* and *ModLoop (157)*.

3. Side-chains are generally modeled by "copying" the conformations of conserved residues from their structural template(s) and selecting those of mutated residues from libraries containing the most common conformations that each residue assumes in protein structures (rotamers). After all non-conserved residues have been replaced, unfavorable van der Waals contacts might be present in the model and have to be eliminated, for example, by exploring side-chain conformations different from those involved in the clashes. Alternative side-chain conformations may be examined also to try and bring potentially interacting residues next to each other, for example, in case hydrogen-bond donor or acceptors or, more rarely, positively or negatively charged groups, found in a buried region of the model do not have an interaction partner nearby.

Several programs have been developed to automatically model side-chain conformations, which take into account the above factors and explore combinatorially a number of rotamers for each residue of the model to try and find an optimal conformational ensemble. *SCRWL (158)* was used by several groups providing good rotamer predictions in the CM category in *CASP6 (14)* where, as expected, prediction of rotamers was found to be more difficult for surface than for buried residues, which are subjected to stronger constraints. However, methods providing the best rotamer predictions were not the best at predicting side-chain contacts, which are, in turn, best predicted at the expense of rotamer accuracy *(14)*. The program *Protinfo AB CM (159)* and the *Scap* program *(160)* of the *Jackal* package were used by some of the best predictors of side-chain contacts in the CM category in *CASP6*.

The ability of several programs (*Modeller (161)*, *SegMod/ENCAD (162)*, *SwissModel (163)*, *3D-JIGSAW (164)*, *Nest (165)* of the *Jackal* package, and Builder (*166*)) to build 3D models starting from target-template sequence alignments has been assessed *(167)*. In this test *Modeller*, *Nest*, and *SegMod/ENCAD* performed better than the others, although no program was better than all the others in all tests. *Modeller* is the program used by most of the successful *CASP* groups to generate 3D models from target-template alignments produced using SC- or FR-based methods. The relative performance of *SwissModel (163)*, *CPHmodels (168)*, *3D-JIGSAW*, and *ESyPred3D (169)* is continuously evaluated by *EVA*.

3.3.4. Model Evaluation

Programs like *Procheck*, *Whatcheck (170)*, and *Swiss-PDB Viewer* evaluate the quality of 3D structures based on parameters such as the number of main-chain dihedral angles lying outside the allowed regions of the Ramachandran Plot, unfavorable van der Waals contacts, and buried polar residues not involved in hydrogen bond formation. Some of these programs also evaluate the

energy of bond lengths, angles, torsions, and electrostatic interactions based on empirical force fields. These evaluations can be useful to highlight errors in the model resulting from the modeling procedure, or more rarely, inherited from the template structures. However, it is worth stressing that the stereochemical correctness of the model is no guarantee of its biological accuracy, i.e., its actual similarity to the target structure.

Other model quality assessment programs (MQAPs), such as *ProSa (171)*, *Verify-3D (15, 172)*, and *ANOLEA (173, 174)* evaluate model quality based on the comparison between the 3D context of each target residue in the model and the 3D context commonly associated with each residue in known structures. Environmental features taken into account by these programs include neighboring residues or atoms, solvent accessible surface area, and secondary structure of each residue. Similar structural features are incorporated in the neural-network based *ProQ (175)* and *ProQres (176)* programs, which assign quality measures to protein models or parts of them, respectively. Other tools, such as *COLORADO3D*, help evaluate model quality by visualizing information such as sequence conservation, solvent accessibility, and potential errors, including those detected by *ProSa*, *Verify-3D*, and *ANOLEA*. *MQAP-Consensus (177)* uses a consensus of MQAP methods registered in *CAFASP* to evaluate and select models produced by different servers, all of which can be downloaded from the *CAFASP4 MQAP* Web server. One or more of these MQAPs were used by the most successful predictors in *CASP6* to evaluate their models at various stages of the model building and refinement procedures. Successful prediction strategies included collecting models from different servers or building a number of models based on different templates and/or different target-template alignments, and screening them based on quality assessments performed by MQAPs. The regions that are structurally conserved in the different models and/or are considered to be reliable by MQAPs are retained, whereas structurally variable and/or less reliable regions according to MQAPs are realigned and remodeled until they either reach an acceptable quality, as measured by MQAP methods, or their quality cannot be improved anymore in subsequent refinement cycles.

Blind predictions of both the overall and residue-based quality of protein models were analyzed by human assessors for the first time in *CASP7*. The relative assessment paper in the 2007 issue of the Journal *PROTEINS* dedicated to CASP7 provides a reliable picture of the relative performance of MQAPs in blind tests.

3.3.5. Model Refinement

Based on the results of quality assessment programs, 3D models can be refined to eliminate obvious structural mistakes (for example by selecting alternative side-chain rotamers to eliminate unfavorable residue–residue interactions) using structure visualization

programs such as *Swiss-PDB Viewer* and *InsightII/Biopolymer (178)*. *Swiss-PDB Viewer*, *GROMACS (179)*, *NAMD (180)*, *TINKER*, and the *Discover* module of *InsightII* can also perform energy minimizations. Since no evidence has been collected over the various *CASP* experiments about model improvements achieved by energy minimization procedures, these should only be limited to the regions showing bad geometrical parameters (e.g., those in which fragments from different structures have been joined, for example, in loop modeling) and involved in clashes that cannot be relieved by changing side-chain conformations (e.g., those involving main-chain atoms and/or proline side-chains). However, as discussed for loop modeling, since there is no evidence that such procedures will improve the model (i.e., make it more similar to the target structure, as opposed to improving its geometry) rather than make it worse, and depending on the proximity of any problematic regions to important structural/functional sections of the model, small structural errors may also be left unrefined and indicated as potentially less reliable regions in the model.

The increase in evolutionary distance between target and template is associated with a reduction of the conserved core and an enlargement of the variable regions, which in turn makes it more and more difficult to align the target and template sequences correctly. For distantly related templates (e.g., those identified by FR methods) errors in the target-template(s) sequence alignment might result in serious mistakes that cannot be corrected by modifying the coordinates of the final model. When this occurs, it is necessary to re-evaluate the target-template alignment by making use of any 3D information contained in the model, and try to modify the alignment in such a way that the new model generated from it will not contain the aforementioned errors. Several cycles of model building and evaluation might be necessary to achieve this result.

3.4. Structure Prediction Without Template

As discussed in the preceding, evolutionarily related domains of known structure are used as templates for the whole target. When no template structure can be identified in the DBs by either sequence-based or FR methods, two scenarios are possible: either a structure similar to the target is present in the DBs, but none of the aforementioned SC- or FR-based methods is able to detect it, or no similar structure is available, i.e., the target structure is actually a NF. In both cases, different methods from those described before have to be used. The so-called *ab initio* methods, which use computationally intensive strategies attempting to recreate the physical and chemical forces involved in protein folding, have been, until now, less successful at predicting protein structures in absence of structural templates than the knowledge-based approaches. These exploit information contained in the

structure databases, from which fragments potentially similar to the targets are extracted and assembled together to produce a full 3D model of the target *(18, 19)*.

Models built by fragment-assembly techniques are evaluated using knowledge-based statistical potentials and clustering procedures. Knowledge-based potentials contain terms derived from statistical analyses of protein structures as well as physicochemical terms (e.g., terms for pairwise interactions between residues or atoms, residue hydrophobicity, hydrogen bonds, and main-chain and side-chain dihedral angles). Since large numbers of conformations are generated for each target, a clustering analysis is performed based on structure similarity and the cluster centroid model is usually chosen. Models representative of highly populated clusters are assumed to be more likely to be correct than models from less populated clusters. Several NF methods take advantage of homologous sequence information, either for secondary structure predictions or as input for model building, and some use 3D models produced by automated CM, FR, and/or NF prediction servers and consensus SSE predictions to derive structural restraints that are used to guide or constrain a subsequent fragment assembly procedure or folding simulation. Manual intervention can occur at different stages, for example, to choose templates or fragments, or inspect models.

Although originally developed for NF predictions, fragment-based methods have been used by several successful predictors in the FR category in both *CASP6* and *CASP5*, and the group that developed the program *Rosetta (19, 182–186)* and the server *Robetta (187–189)* has been the most successful at predicting the structure of difficult FR targets, i.e., targets for which no evolutionary relationship with known structures is apparent and are classified in the *CASP* FR/A (A: analogous) sub-category *(190, 191)*.

Unfortunately, the performance of these methods on real NF targets is somewhat less successful. In *CASP6*, nine targets whose structures did not show overall similarities with already known folds based on the results of the *LGA* program *(68)* were evaluated in the NF category *(190)*. Three of them, however, turned out to be variants of known folds, in that they contain sub-structures that match sub-structures in known proteins, and were therefore defined as NF "easy" *(190)*. In the NF category, models bearing an overall structural similarity with the target structures were submitted only for these three "easy" targets. For the remaining six NF "hard" targets, which did not show any similarity to known folds (and were, therefore, the only truly "novel" folds), no globally correct prediction was submitted. It should be mentioned, however, that all of the NF "hard" targets were relatively large proteins (*115–213* a.a.s vs. *73–90* a.a.s of NF "easy" targets), which are particularly difficult to predict by NF methods. The best structure predictions for these targets were

limited to partial regions, although often larger than standard supersecondary structures, which could not be assembled into correct global topologies. Since it is not obvious how to compare predictions for different target fragments in the context of globally incorrect models, it is particularly difficult to evaluate these predictions and decide which of the methods were the most successful. Further, the difference between the best and average models submitted for these targets is not very large, and since most models are poor even large differences in ranking are unlikely to be significant. Another general drawback of methods used to predict NF targets is that predictors were not particularly good at recognizing their best models, which highlights a problem with ranking. Finally, even considering all the nine NF targets, the sample size is too small to draw statistically significant conclusions *(190)*. Taking all these caveats into account, FRAGFOLD *(192)*, CABS *(193)*, and *Rosetta* were used by the best performing groups in the NF category, although several other groups submitted each at least one best model for different targets. Another freely available program that was used by a relatively well-performing group is *ROKKY* (194).

The *Rosetta* folding simulation program by the Baker group *(19, 182–184)* is not only one of the best available methods for NF and difficult FR targets predictions (the Baker group was the best or among the best performing at predictions without template in the last *CASP* editions) but, being freely available, it is also one of the most popular and one that has been incorporated in a large number of metaservers (*see* **Section 3.5.2**). The output of *Rosetta* contains a number of 3D models of the target protein ordered according to an energy score. As for FR methods, this score cannot be solely relied upon, and all additional sequence, structural, functional, and evolutionary information that can be gathered about the target should be exploited. Since a sequence for which no prediction can be provided by either SC- or FR-methods might be a difficult FR target (FR/A) rather than an NF, top scoring models and other models that might be correct based on all available information about the target should be used to search structural DBs with structural alignment programs (e.g., *DALI (31)*, *VAST*, *CE*, *SSM*, *Mammoth-mult*). If proteins of known structure showing significant structural similarity with these models can be found (i.e., the sequence is likely to be a difficult FR target), these should be examined to identify sequence, structural, and/or functional evidence (e.g., key residues, *see* **Section 3.3.2.5**) that might support the hypothesis of the correctness of the model(s) produced for the target. Conversely, if no protein showing global structural similarity with the target is detected, the target might indeed have a novel fold. Nevertheless, searches of structural DBs might detect proteins containing sub-structures similar to those of the target, and the analysis of

these structurally similar regions might provide important clues supporting some models over others.

Given their independence from structural templates, NF techniques can be useful to try to predict not only complete 3D structures for targets that elude CM and FR methods, but also loops and other regions of models generated by template-based methods that are not conserved with respect to the available templates.

3.5. Automated Methods, Metaservers, and 3D Model DBs

3.5.1. Automated Methods

A number of automated CM, FR, and NF methods and metaservers for protein structure prediction are described in the relevant sections of this chapter (e.g., *see* **Sections 3.3.1.1, 3.3.1.2, 3.3.3.2, 3.4, and 3.5.2**), together with their performance in the hands of expert users. As far as the performance of automated methods without user intervention is concerned, *SPARKS2* and *SP3* have been the first and third best performing servers in the CM category in *CASP6*, and the 14th and 22nd overall best predictors in ranking including human predictors *(14)*. The second best server was *STRUCTFAST (195)*. In the FR category *Robetta* was the best server and the overall ninth best predictor in ranking including human predictors. In the NF category, *Robetta* was the best server. The best performing servers in *Livebench* and *CAFASP* are all metaservers (*see* **Section 3.5.2**), in particular *3D-Jury* (196), *3D-Shotgun (197–200)*, *Pcons*, and *Pmodeller (201)*.

In spite of the good performance of the best automated methods in blind tests, intervention by expert users (often by the same authors of the automated procedures) consistently provides significantly improved results with respect to fully automated methods, indicating that, in spite of much effort, the prediction community has not yet managed to encode all expert knowledge on protein structure prediction into automated programs. Therefore, depending on the number of models required, the purpose for which they are built, and the user's level of expertise, automated programs might be best used as a support for, rather than a replacement of, human predictions.

3.5.2. Metaservers

Metaservers collect a variety of predictions and structural information from different automated servers, evaluate their quality, and combine them to generate consensus predictions that might consist either in the result provided by a single structure prediction server and reputed to be the best among those examined, or in the combination of results provided by different servers. Putative *PDB* templates, target-template sequence alignments, and 3D structural models are collected from fully automated SC, FR, or NF method-based servers and in some cases other metaservers that perform regularly well in blind test experiments such as *CASP*, *CAFASP*, or *Livebench*. This information is often integrated with domain, secondary structure, and function predictions,

as well as with information derived from analyses of protein structures, e.g., protein structure classification by *SCOP* and/or *FSSP* and secondary structure assignment by *DSSP*. Depending on the metaserver, different types of predictions are returned, consisting of 3D model coordinates *(202)* or target-template sequence alignments from which 3D models can then be generated *(113)*. Preliminary models may be fragmented and spliced to produce new hybrid models and/or subjected to refinements to eliminate errors such as overlapping regions or broken chains. Predictions of domain boundaries, secondary structure, transmembrane helices, and disordered regions are often provided as well.

Metaservers exploit a variety of criteria and methods to evaluate the quality of alignments and models that they collect and produce, including: the scores associated with the predictions by contributing servers; the extent to which different alignments agree with one another; contact order, i.e., the average sequence separation between all pairs of contacting residues normalized by the total sequence length *(203, 204)*; structural comparisons of independent models and detection of large clusters of structures with similar conformations *(205)*; and model/structure evaluation programs mentioned before, such as *ProSa*, *Verify-3D*, *ProQ*, *ProQres*, and *COLORADO3D*. Well-scoring model fragments (e.g., consensus predictions provided by unrelated methods) are considered to be reliable and are included in the model, whereas for poorly scoring regions models based on alternative alignments are generated and assessed using MQAPs until they reach acceptable quality or their quality cannot be further improved.

Many of the available metaservers have implemented completely automated procedures to cover the full spectrum of structure predictions, from CM to NF, by combining template-based and *de novo* structure prediction methods. First, several well-performing CM and FR servers are queried with the sequence of the target protein. If homologs with experimentally characterized 3D structure are detected, the conserved regions of the target are predicted by template-based modeling and the variable regions (e.g., loops, N- and C-terminal extensions) by NF methods such as the *Rosetta* fragment-insertion protocol. If no reliable structural homolog is found, the target sequence is sent to NF prediction servers (most often *Robetta*).

Predictors that used metaservers performed among the best in the more recent *CASP* editions, in both CM *(14)* and FR *(191)* categories. Although metaservers owe much of their predictive power to the quality of the results provided by remote servers, their ability to choose, evaluate, and combine different predictions has often resulted in better performances than those achieved by the remote prediction servers alone.

Meta-BASIC and *3D-Jury* were used by the top ranking predictor in both *CASP6* CM and FR/H categories *(14, 191)*, and

3D-Jury was also used by another of the top four predictors in the CM category *(206)*.

The group that developed the *Robetta* server ranked best in FR/A, among the best in FR and NF, and well in the CM category in *CASP6*. The *Robetta* server itself performed well in both the CM and FR categories in *CASP6*.

The *Genesilico* metaserver *(202)* was used by one of the most successful groups in the CM and FR categories in *CASP5* and was one of the best in all categories in *CASP6*, where *Genesilico* was used in combination with the *FRankenstein3D* program for splicing of FR models *(207)* and fragment assembly techniques for NF targets *(193)*.

Pcons, *Pmodeller* and *Pcons5 (208, 209)*, and *Bioinbgu*, *3D-Shotgun* and *Inub (197–200)* were among the best performing servers in *CASP5* and groups using them performed well, although not among the best, in both the CM and FR categories in *CASP6*.

The groups using *@TOME (210)* and a combination of CHIMERA and FAMS *(211)* performed well, although they were not among the best groups, in *CASP6*.

3.5.3. 3D Model DBs

Before embarking on the modeling procedure, however automatically, it might be worth checking whether 3D models for the target of interest are available among protein model DBs. The *SwissModel Repository (212)* and *ModBase (213)* contain 3D models of all sequences in the *SwissProt* and *TrEMBL* databases that have detectable sequence similarity with proteins of known structure, generated by the *SwissModel* and *Modeller* programs, respectively. *FAMSBASE (214)*, built using the fully automated modeling system FAMS, contains comparative models for all the proteins assigned to many sequenced genomes. Of course, models produced automatically, even those for "easy" CM targets, may contain significant errors; to detect them, models should be subjected to all the sequence, structure, function, and evolutionary analyses described before. The Protein Model Database (*PMDB*) *(215)* stores manually built 3D models together with any supporting information provided. It contains, among others, the models submitted to the past *CASP* editions, and it allows users to submit, as well as to retrieve, 3D protein models.

3.6. Expected Accuracy of the Model

A general indication of the accuracy of template-based models is provided by sequence–structure relationships derived from the analysis of homologous proteins *(13)*. Based on these analyses, if the sequence identity between target and template in the conserved core regions is about 50% or higher, we can expect the main-chain atoms of the core regions of the target structure to be superimposable to the best template and, consequently, to the model built on this template, with an RMSD value (*see* **Note 1**) within 1.0 Å. However, for sequence identities as low as 20%,

the structural similarity between two proteins can vary to a large extent *(13)*.

Similar results are provided by the structural comparison of the target domain structures and the best models produced for them in *CASP6* (**Fig. 2.2**). For %_ID values between the sequences of target and best-template of 50% or higher, the GDT_TS (*see* **Note 1**) between the real target structure and the best model submitted for it was 90 or higher. For %_ID values between 20% and 30% the GDT_TS varied between about 100 and 50. For %_ID values <20%, the GDT_TS was between 90 and 40 for CM targets, between 90 and 30 for FR targets, and from 60 to <20 for NF targets. In other words, for sequence identities of 20% or lower, the model built can vary from either having a globally very similar structure to the target (GDT_TS >80) to only having a few small fragments showing some structural similarity to the target (GDT_TS <20). A visual example of structural variations corresponding to representative GDT_TS values is shown in **Fig. 2.2**.

In *CASP6*, for "easy" CM targets template detection and production of accurate target-template sequence alignments were carried out successfully and the accuracy of the produced models was quite high, at least in the conserved regions. Conversely, model refinement and improvement over the best template were still open issues: In only a few cases were the best predictors able to produce 3D models more similar to their target structures than the structures of the best templates. In most cases, improvements over the best templates were obtained only for the easiest targets, whereas for harder targets the best templates were generally much more similar to the target structures than any model. In *CASP7*, for the first time, models of such "easy" CM targets were assessed in a separate category to establish whether there has been any improvement in these areas. For "hard" *CASP6* CM targets choosing the right template and producing the correct target-template alignment proved to be challenging tasks *(14)*, and both became more and more difficult with an increase in the evolutionary distance between targets and best templates. For most FR targets the top predictions contained all or most of the SSE in the right place or with small shifts with respect to the target structure. However, for many targets most predictions had little resemblance to the target structure *(191)*. For "hard" NF targets, only fragments of the best models resembled the real target structures *(190)*.

Structural comparisons of the templates selected for a specific target (*see* **Section 3.3.2.4**) permit identification of relationships between %_ID and structural similarity for the target family, based on which a more accurate prediction about the expected model accuracy for targets belonging to that family can be obtained. In any case, information about model accuracy provided by sequence–structure relationships is limited to the

conserved regions between target and template(s). Unless loop regions belong to one of the restricted number of categories for which sequence–structure relationships have been defined (*see* **Section 3.3.3.2**), or side-chain conformations are strongly conserved for structural and/or functional reasons, in general it is not possible to know how accurate the predictions of loops or side-chain conformations are before comparing them with the real structures.

Automated programs can generally produce models of "easy" CM targets (e.g., sequence identity with the templates >40%) that are good enough for many practical applications. However, these models may contain serious mistakes even in regions close to the functional sites, the more so the higher the evolutionary distance between target and template; therefore, it is wise to carefully compare their output with the template structures, especially in the key structural and functional regions. This is a point of utmost importance: In principle, it is always possible to build homology-based 3D models automatically; all that is needed is a template structure and an alignment between the target and template sequences. However, if the chosen template is not a real target homologue, or if the target-template sequence alignment is wrong, the model is bound to be incorrect ("garbage *in*, garbage *out*!"). In *CASP*, expert human predictors using semi-automated procedures consistently produce better predictions than fully automated methods. However, in choosing between the use of a fully automated method and manual intervention it is important to remember that in both CM and FR categories there was a big difference between the best predictors and many of the others, and that the best automated programs (*see* **Section 3.5.1**) performed remarkably well compared with many human predictors.

4. Notes

1. The two commonly used measures to evaluate similarity between protein structures used in this chapter are root-mean-square deviation (RMSD) and global distance test total score (GDT_TS). The RMSD between two structures A and B is the average distance between a specific set of atoms in structure A (e.g., Cα or main-chain atoms belonging to a given set of residues) and the corresponding atoms in structure B, after optimal superposition of their coordinates. The formula to calculate the RMSD is:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}}$$

where d_i is the distance between the pair of atoms i and N is the total number of superimposed atom pairs in the dataset. The GDT_TS between two structures A and B represents the percentage of residues of protein A whose Cα atoms are found within specific cut-off distances from the corresponding Cα atoms in structure B after optimal superimposition of the two structures. In *CASP* the GDT_TS is calculated according to the formula:

$$\text{GDT\_TS} = (\text{GDT\_P1} + \text{GDT\_P2} + \text{GDT\_P4} + \text{GDT\_P8})/4,$$

where GDT_P1, GDT_P2, GDT_P4, and GDT_P8 represent the number of Cα atoms of a target model whose distance from the corresponding atoms of the real structure, after optimal model-to-structure superposition, is ≤1, 2, 4, and 8 Å, respectively, divided by the number of residues in the target structure.

2. The score of the *BLAST* alignment is calculated as the sum of substitution and gap scores. Substitution scores are given by the chosen substitution matrix (e.g., PAM, BLOSUM, *see* **Note 3**), whereas gap scores are calculated as the sum of the gap opening and gap extension penalties used in the search. The choice of these penalties is empirical, but in general a high value is chosen for gap opening and a low value for gap extensions (e.g., *BLAST* default values for gap insertions and extensions are 7–12 and 1–2, respectively). The rational bases for these choices are that: (1) only a few regions of protein structures (e.g., loops) can accommodate insertions and deletions (hence, a high value for gap openings); and (2) most of these regions can usually accept insertions and deletions of more than a single residue (hence, the lower value used for gap extension).
3. Substitution matrices are 20 × 20 matrices containing a value for each a.a. residue pair that is proportional to the probability that the two a.a.s substitute for each other in alignments of homologous proteins. The two most commonly used matrix series, PAM (percent accepted mutation) *(216)* and BLOSUM (blocks substitution matrix) *(217)*, comprise several matrices that have been designed to align proteins with varying extent of evolutionary distance. PAM matrices with high numbers (e.g., PAM 250) and BLOSUM matrices with low numbers (e.g., BLOSUM 45) are suitable for aligning distantly related sequences; conversely, low PAM (e.g., PAM 1) and high BLOSUM (e.g., BLOSUM 80) number matrices are appropriate to align closely related sequences. *BLAST* default matrix is BLOSUM 62.
4. In *SCOP*, protein domains are assigned to the same family if they either have %_ID ≥ 30% or, in case their %_ID is lower,

if they have very similar structures and functions. Different families are assigned to the same superfamily if they comprise proteins presenting structural and, often, functional features suggesting a common evolutionary origin. Families and superfamilies having the same major SSE in the same spatial arrangement and with the same topological connections are assigned to the same fold. Proteins assigned to the same families and superfamilies are thought to be close and remote evolutionary relatives, respectively, whereas proteins having the same fold might still have a common origin but no evidences strong enough to support this hypothesis are available yet.

Acknowledgments

The authors gratefully acknowledge Claudia Bertonati, Gianni Colotti, Andrea Ilari, Romina Oliva, and Christine Vogel for manuscript reading and suggestions, and Julian Gough and Martin Madera for discussions.

References

1. Moult, J., Pedersen, J. T., Judson, R., et al. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins* 23, ii–v.
2. Moult, J., Hubbard, T., Bryant, S. H., et al. (1997) Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins* Suppl. 1, 2–6.
3. Moult, J., Hubbard, T., Fidelis, K., et al. (1999) Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins* Suppl. 3, 2–6.
4. Moult, J., Fidelis, K., Zemla, A., et al. (2001) Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins* Suppl. 5, 2–7.
5. Moult, J., Fidelis, K., Zemla, A., et al. (2003) Critical assessment of methods of protein structure prediction (CASP): round V. *Proteins* 53, Suppl. 6, 334–339.
6. Moult, J., Fidelis, K., Rost, B., et al. (2005) Critical assessment of methods of protein structure prediction (CASP): round 6. *Proteins* 61, Suppl. 7, 3–7.
7. Fischer, D., Barret, C., Bryson, K., et al. (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins* Suppl. 3, 209–217.
8. Fischer, D., Elofsson, A., Rychlewski, L., et al. (2001) CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins* Suppl. 5, 171–183.
9. Fischer, D., Rychlewski, L., Dunbrack, R. L., Jr., et al. (2003) CAFASP3: the third critical assessment of fully automated structure prediction methods. *Proteins* 53, Suppl. 6, 503–516.
10. Berman, H. M., Westbrook, J., Feng, Z., et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28, 235–242.
11. Rychlewski, L., Fischer, D. (2005) Live Bench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci* 14, 240–245.
12. Koh, I. Y., Eyrich, V. A., Marti-Renom, M. A., et al. (2003) EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res* 31, 3311–3315.
13. Chothia, C., Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *Embo J* 5, 823–826.
14. Tress, M., Ezkurdia, I., Grana, O., et al. (2005) Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins* 61, Suppl. 7, 27–45.

15. Bowie, J. U., Luthy, R., Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253, 164–170.
16. Jones, D. T., Taylor, W. R., Thornton, J. M. (1992) A new approach to protein fold recognition. *Nature* 358, 86–89.
17. Sippl, M. J., Weitckus, S. (1992) Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins* 13, 258–271.
18. Jones, D. T. (1997) Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins* Suppl. 1, 185–191.
19. Simons, K. T., Kooperberg, C., Huang, E., et al. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 268, 209–225.
20. Sprague, E. R., Wang, C., Baker, D., et al. (2006) Crystal structure of the HSV-1 Fc receptor bound to Fc reveals a mechanism for antibody bipolar bridging. *PLoS Biol* 4, e148.
21. Galperin, M. Y. (2006) The Molecular Biology Database Collection: 2006 update. *Nucleic Acids Res* 34, D3–5.
22. Fox, J. A., McMillan, S., Ouellette, B. F. (2006) A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Res* 34, W3–5.
23. Benson, D. A., Boguski, M. S., Lipman, D. J., et al. (1997) GenBank. *Nucleic Acids Res* 25, 1–6.
24. Wu, C. H., Apweiler, R., Bairoch, A., et al. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* 34, D187–191.
25. Coutinho, P. M., Henrissat, B. (1999) Carbohydrate-active enzymes: an integrated database approach. In *Recent Advances in Carbohydrate Bioengineering*. H.J. Gilbert, G. Davies, B. Henrissat and B. Svensson eds., The Royal Society of Chemistry, Cambridge, UK, pp. 3–12.
26. Lander, E. S., Linton, L. M., Birren, B., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
27. LoVerde, P. T., Hirai, H., Merrick, J. M., et al. (2004) Schistosoma mansoni genome project: an update. *Parasitol Int* 53, 183–192.
28. Andreeva, A., Howorth, D., Brenner, S. E., et al. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 32, D226–229.
29. Pearl, F., Todd, A., Sillitoe, I., et al. (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res* 33, D247–251.
30. Holm, L., Ouzounis, C., Sander, C., et al. (1992) A database of protein structure families with common folding motifs. *Protein Sci* 1, 1691–1698.
31. Holm, L., Sander, C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res* 25, 231–234.
32. Mizuguchi, K., Deane, C. M., Blundell, T. L., et al. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci* 7, 2469–2471.
33. Gasteiger, E., Gattiker, A., Hoogland, C., et al. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* 31, 3784–3788.
34. Smith, R. F., Wiese, B. A., Wojzynski, M. K., et al. (1996) BCM Search Launcher—an integrated interface to molecular biology data base search and analysis services available on the World Wide Web. *Genome Res* 6, 454–462.
35. Stothard, P. (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* 28, 1102, 1104.
36. Janin, J. (2005) Assessing predictions of protein-protein interaction: the CAPRI experiment. *Protein Sci* 14, 278–283.
37. Janin, J., Henrick, K., Moult, J., et al. (2003) CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* 52, 2–9.
38. Tai, C. H., Lee, W. J., Vincent, J. J., et al. (2005) Evaluation of domain prediction in CASP6. *Proteins* 61, Suppl. 7, 183–192.
39. Kim, D. E., Chivian, D., Malmstrom, L., et al. (2005) Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. *Proteins* 61, Suppl. 7, 193–200.
40. Suyama, M., Ohara, O. (2003) DomCut: prediction of inter-domain linker regions in amino acid sequences. *Bioinformatics* 19, 673–674.
41. Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., et al. (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res* 33, D192–196.

42. Finn, R. D., Mistry, J., Schuster-Bockler, B., et al. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* 34, D247–251.
43. Letunic, I., Copley, R. R., Pils, B., et al. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* 34, D257–260.
44. Bru, C., Courcelle, E., Carrere, S., et al. (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res* 33, D212–215.
45. Mulder, N. J., Apweiler, R., Attwood, T. K., et al. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res* 33, D201–205.
46. Hulo, N., Bairoch, A., Bulliard, V., et al. (2006) The PROSITE database. *Nucleic Acids Res* 34, D227–230.
47. Gough, J., Chothia, C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res* 30, 268–272.
48. Madera, M., Vogel, C., Kummerfeld, S. K., et al. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res* 32, D235–239.
49. Jin, Y., Dunbrack, R. L., Jr. (2005) Assessment of disorder predictions in CASP6. *Proteins* 61, Suppl. 7, 167–175.
50. Obradovic, Z., Peng, K., Vucetic, S., et al. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* 61, Suppl. 7, 176–182.
51. Peng, K., Radivojac, P., Vucetic, S., et al. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 7, 208.
52. Cheng, J., Sweredoski, M., Baldi, P. (2005) Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining Knowl Disc* 11, 213–222.
53. Dosztanyi, Z., Csizmok, V., Tompa, P., et al. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21, 3433–3434.
54. Vullo, A., Bortolami, O., Pollastri, G., et al. (2006) Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res* 34, W164–168.
55. Ward, J. J., Sodhi, J. S., McGuffin, L. J., et al. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337, 635–645.
56. Bryson, K., McGuffin, L. J., Marsden, R. L., et al. (2005) Protein structure prediction servers at University College London. *Nucleic Acids Res* 33, W36–38.
57. Krogh, A., Larsson, B., von Heijne, G., et al. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305, 567–580.
58. Rost, B., Yachdav, G., Liu, J. (2004) The PredictProtein server. *Nucleic Acids Res* 32, W321–326.
59. Bagos, P. G., Liakopoulos, T. D., Spyropoulos, I. C., et al. (2004) PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. *Nucleic Acids Res.* 32, W400–404.
60. Natt, N. K., Kaur, H., Raghava, G. P. (2004) Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins* 56, 11–18.
61. Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 292, 195–202.
62. Karplus, K., Barrett, C., Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14, 846–856.
63. Pollastri, G., McLysaght, A. (2005) Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics* 21, 1719–1720.
64. Cuff, J. A., Barton, G. J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 40, 502–511.
65. Cuff, J. A., Clamp, M. E., Siddiqui, A. S., et al. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics* 14, 892–893.
66. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389–3402.
67. Tress, M., Tai, C. H., Wang, G., et al. (2005) Domain definition and target classification for CASP6. *Proteins* 61, Suppl. 7, 8–18.
68. Pearson, W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol* 183, 63–98.
69. Pearson, W. R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci* 4, 1145–1160.
70. Park, J., Karplus, K., Barrett, C., et al. (1998) Sequence comparisons using

multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* 284, 1201–1210.

71. Eddy, S. R. (1996) Hidden Markov models. *Curr Opin Struct Biol* 6, 361–365.
72. Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics* 14, 755–763.
73. Madera, M., Gough, J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res* 30, 4321–4328.
74. Karplus, K., Karchin, R., Draper, J., et al. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* 53, Suppl. 6, 491–496.
75. Karplus, K., Katzman, S., Shackleford, G., et al. (2005) SAM-T04: what is new in protein-structure prediction for CASP6. *Proteins* 61, Suppl. 7, 135–142.
76. Schaffer, A. A., Wolf, Y. I., Ponting, C. P., et al. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* 15, 1000–1011.
77. Ohlson, T., Wallner, B., Elofsson, A. (2004) Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. *Proteins* 57, 188–197.
78. Yona, G., Levitt, M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J Mol Biol* 315, 1257–1275.
79. von Ohsen, N., Sommer, I., Zimmer, R. (2003) Profile-profile alignment: a powerful tool for protein structure prediction. *Pac Symp Biocomput* 252–263.
80. von Ohsen, N., Sommer, I., Zimmer, R., et al. (2004) Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. *Bioinformatics* 20, 2228–2235.
81. Sadreyev, R., Grishin, N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol* 326, 317–336.
82. Mittelman, D., Sadreyev, R., Grishin, N. (2003) Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. *Bioinformatics* 19, 1531–1539.
83. Sadreyev, R. I., Baker, D., Grishin, N. V. (2003) Profile-profile comparisons by COMPASS predict intricate homologies between protein families. *Protein Sci* 12, 2262–2272.
84. Heger, A., Holm, L. (2001) Picasso: generating a covering set of protein family profiles. *Bioinformatics* 17, 272–279.
85. Edgar, R. C., Sjolander, K. (2004) COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics* 20, 1309–1318.
86. Pietrokovski, S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res* 24, 3836–3845.
87. Jaroszewski, L., Rychlewski, L., Li, Z., et al. (2005) FFAS03: a server for profile–profile sequence alignments. *Nucleic Acids Res* 33, W284–288.
88. Tomii, K., Akiyama, Y. (2004) FORTE: a profile-profile comparison tool for protein fold recognition. *Bioinformatics* 20, 594–595.
89. Ginalski, K., Pas, J., Wyrwicz, L. S., et al. (2003) ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Res* 31, 3804–3807.
90. Soding, J., Biegert, A., Lupas, A. N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33, W244–248.
91. Kabsch, W., Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637.
92. Sippl, M. J. (1995) Knowledge-based potentials for proteins. *Curr Opin Struct Biol* 5, 229–235.
93. Kelley, L. A., MacCallum, R. M., Sternberg, M. J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* 299, 499–520.
94. Jones, D. T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* 287, 797–815.
95. McGuffin, L. J., Bryson, K., Jones, D. T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics* 16, 404–405.
96. Zhang, Y., Arakaki, A. K., Skolnick, J. (2005) TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins* 61, Suppl. 7, 91–98.
97. Skolnick, J., Kihara, D., Zhang, Y. (2004) Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins* 56, 502–518.
98. Shi, J., Blundell, T. L., Mizuguchi, K. (2001) FUGUE: sequence-structure homology

recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* 310, 243–257.

99. Xu, J., Li, M., Kim, D., et al. (2003) RAPTOR: optimal protein threading by linear programming. *J Bioinform Comput Biol* 1, 95–117.
100. Tang, C. L., Xie, L., Koh, I. Y., et al. (2003) On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *J Mol Biol* 334, 1043–1062.
101. Teodorescu, O., Galor, T., Pillardy, J., et al. (2004) Enriching the sequence substitution matrix by structural information. *Proteins* 54, 41–48.
102. Zhou, H., Zhou, Y. (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* 55, 1005–1013.
103. Zhou, H., Zhou, Y. (2005) SPARKS 2 and SP3 servers in CASP6. *Proteins* 61, Suppl. 7, 152–156.
104. Zhou, H., Zhou, Y. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* 58, 321–328.
105. Thompson, J. D., Higgins, D. G., Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673–4680.
106. Notredame, C., Higgins, D. G., Heringa, J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302, 205–217.
107. Thompson, J. D., Gibson, T. J., Plewniak, F., et al. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25, 4876–4882.
108. Crooks, G. E., Hon, G., Chandonia, J. M., et al. (2004) WebLogo: a sequence logo generator. *Genome Res* 14, 1188–1190.
109. Sonnhammer, E. L., Hollich, V. (2005) Scoredist: a simple and robust protein sequence distance estimator. *BMC Bioinformatics* 6, 108.
110. Galtier, N., Gouy, M., Gautier, C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* 12, 543–548.
111. Parry-Smith, D. J., Payne, A. W., Michie, A. D., et al. (1998) CINEMA—a novel colour INteractive editor for multiple alignments. *Gene* 221, GC57–63.
112. Ginalski, K., von Grotthuss, M., Grishin, N. V., et al. (2004) Detecting distant homology with Meta-BASIC. *Nucleic Acids Res* 32, W576–581.
113. Xu, Y., Xu, D., Gabow, H. N. (2000) Protein domain decomposition using a graph-theoretic approach. *Bioinformatics* 16, 1091–1104.
114. Guo, J. T., Xu, D., Kim, D., et al. (2003) Improving the performance of DomainParser for structural domain partition using neural network. *Nucleic Acids Res* 31, 944–952.
115. Alexandrov, N., Shindyalov, I. (2003) PDP: protein domain parser. *Bioinformatics* 19, 429–430.
116. Todd, A. E., Orengo, C. A., Thornton, J. M. (1999) DOMPLOT: a program to generate schematic diagrams of the structural domain organization within proteins, annotated by ligand contacts. *Protein Eng* 12, 375–379.
117. Zemla, A. (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res* 31, 3370–3374.
118. Holm, L., Park, J. (2000) DaliLite workbench for protein structure comparison. *Bioinformatics* 16, 566–567.
119. Ortiz, A. R., Strauss, C. E., Olmea, O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci* 11, 2606–2621.
120. Gibrat, J. F., Madej, T., Bryant, S. H. (1996) Surprising similarities in structure comparison. *Curr Opin Struct Biol* 6, 377–385.
121. Shindyalov, I. N., Bourne, P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11, 739–747.
122. Orengo, C. A., Taylor, W. R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol* 266, 617–635.
123. Krissinel, E., Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr* 60, 2256–2268.
124. Yang, A. S., Honig, B. (1999) Sequence to structure alignment in comparative modeling using PrISM. *Proteins* Suppl. 3, 66–72.

125. Lupyan, D., Leo-Macias, A., Ortiz, A. R. (2005) A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics* 21, 3255–3263.
126. Ye, Y., Godzik, A. (2005) Multiple flexible structure alignment using partial order graphs. *Bioinformatics* 21, 2362–2369.
127. Hill, E. E., Morea, V., Chothia, C. (2002) Sequence conservation in families whose members have little or no sequence similarity: the four-helical cytokines and cytochromes. *J Mol Biol* 322, 205–233.
128. Chothia, C., Jones, E. Y. (1997) The molecular structure of cell adhesion molecules. *Annu Rev Biochem* 66, 823–862.
129. Hill, E., Broadbent, I. D., Chothia, C., et al. (2001) Cadherin superfamily proteins in *Caenorhabditis elegans* and *Drosophila melanogaster*. *J Mol Biol* 305, 1011–1024.
130. Chothia, C., Lesk, A. M. (1987) Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* 196, 901–917.
131. Chothia, C., Lesk, A. M., Tramontano, A., et al. (1989) Conformations of immunoglobulin hypervariable regions. *Nature* 342, 877–883.
132. Al-Lazikani, B., Lesk, A. M., Chothia, C. (1997) Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* 273, 927–948.
133. Morea, V., Tramontano, A., Rustici, M., et al. (1998) Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J Mol Biol* 275, 269–294.
134. Mizuguchi, K., Deane, C. M., Blundell, T. L., et al. (1998) JOY: protein sequence-structure representation and analysis. *Bioinformatics* 14, 617–623.
135. Hubbard, S. J., Thornton, J. M., (1993) NACCESS. Department of Biochemistry and Molecular Biology, University College London.
136. McDonald, I. K., Thornton, J. M. (1994) Satisfying hydrogen bonding potential in proteins. *J Mol Biol* 238, 777–793.
137. Morris, A. L., MacArthur, M. W., Hutchinson, E. G., et al. (1992) Stereochemical quality of protein structure coordinates. *Proteins* 12, 345–364.
138. Laskowski, R. A., MacArthur, M. W., Moss, D. S., et al. (1993) PROCHECK: a program to check the stereochemical quality of protein structures *J Appl Cryst* 26, 283–291.
139. Wallace, A. C., Laskowski, R. A., Thornton, J. M. (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng* 8, 127–134.
140. Laskowski, R. A., Hutchinson, E. G., Michie, A. D., et al. (1997) PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends Biochem Sci* 22, 488–490.
141. Sasin, J. M., Bujnicki, J. M. (2004) COLORADO3D, a web server for the visual analysis of protein structures. *Nucleic Acids Res* 32, W586–589.
142. Landau, M., Mayrose, I., Rosenberg, Y., et al. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* 33, W299–302.
143. Guex, N., Peitsch, M. C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18, 2714–2723.
144. Sayle, R. A., Milner-White, E. J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem Sci* 20, 374.
145. Martz, E. (2002) Protein Explorer: easy yet powerful macromolecular visualization. *Trends Biochem Sci* 27, 107–109.
146. Wang, Y., Geer, L. Y., Chappey, C., et al. (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem Sci* 25, 300–302.
147. Vriend, G. (1990) WHAT IF: a molecular modeling and drug design program. *J Mol Graph* 8, 52–56.
148. Koradi, R., Billeter, M., Wuthrich, K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* 14, 51–55, 29–32.
149. Humphrey, W., Dalke, A., Schulten, K. (1996) VMD: visual molecular dynamics. *J Mol Graph* 14, 33–38, 27–38.
150. Tramontano, A., Chothia, C., Lesk, A. M. (1990) Framework residue 71 is a major determinant of the position and conformation of the second hypervariable region in the VH domains of immunoglobulins. *J Mol Biol* 215, 175–182.
151. Sibanda, B. L., Thornton, J. M. (1985) Beta-hairpin families in globular proteins. *Nature* 316, 170–174.
152. Sibanda, B. L., Blundell, T. L., Thornton, J. M. (1989) Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J Mol Biol* 206, 759–777.
153. Bruccoleri, R. E. (2000) Ab initio loop modeling and its application to homology modeling. *Methods Mol Biol* 143, 247–264.

154. Xiang, Z., Soto, C. S., Honig, B. (2002) Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc Natl Acad Sci U S A* 99, 7432–7437.
155. Tosatto, S. C., Bindewald, E., Hesser, J., et al. (2002) A divide and conquer approach to fast loop modeling. *Protein Eng* 15, 279–286.
156. Fiser, A., Sali, A. (2003) ModLoop: automated modeling of loops in protein structures. *Bioinformatics* 19, 2500–2501.
157. Canutescu, A. A., Shelenkov, A. A., Dunbrack, R. L., Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* 12, 2001–2014.
158. Hung, L. H., Ngan, S. C., Liu, T., et al. (2005) PROTINFO: new algorithms for enhanced protein structure predictions. *Nucleic Acids Res* 33, W77–80.
159. Xiang, Z., Honig, B. (2001) Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol* 311, 421–430.
160. Marti-Renom, M. A., Stuart, A. C., Fiser, A., et al. (2000) Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 29, 291–325.
161. Levitt, M. (1992) Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol* 226, 507–533.
162. Schwede, T., Kopp, J., Guex, N., et al. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res* 31, 3381–3385.
163. Bates, P. A., Kelley, L. A., MacCallum, R. M., et al. (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins* Suppl. 5, 39–46.
164. Petrey, D., Xiang, Z., Tang, C. L., et al. (2003) Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins* 53, Suppl. 6, 430–435.
165. Koehl, P., Delarue, M. (1994) Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol* 239, 249–275.
166. Wallner, B., Elofsson, A. (2005) All are not equal: a benchmark of different homology modeling programs. *Protein Sci* 14, 1315–1327.
167. Lund, O., Frimand, K., Gorodkin, J., et al. (1997) Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng* 10, 1241–1248.
168. Lambert, C., Leonard, N., De Bolle, X., et al. (2002) ESyPred3D: prediction of proteins 3D structures. *Bioinformatics* 18, 1250–1256.
169. Hooft, R. W., Vriend, G., Sander, C., et al. (1996) Errors in protein structures. *Nature* 381, 272.
170. Sippl, M. J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins* 17, 355–362.
171. Luthy, R., Bowie, J. U., Eisenberg, D. (1992) Assessment of protein models with three-dimensional profiles. *Nature* 356, 83–85.
172. Melo, F., Devos, D., Depiereux, E., et al. (1997) ANOLEA: a www server to assess protein structures. *Proc Int Conf Intell Syst Mol Biol* 5, 187–190.
173. Melo, F., Feytmans, E. (1998) Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol* 277, 1141–1152.
174. Wallner, B., Elofsson, A. (2003) Can correct protein models be identified? *Protein Sci* 12, 1073–1086.
175. Wallner, B., Elofsson, A. (2006) Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci* 15, 900–913.
176. Fischer, D. (2006) Servers for protein structure prediction. *Current Opin Struct Biol* 16, 178–182.
177. Dayringer, H. E., Tramontano, A., Sprang, S. R., et al. (1986) Interactive program for visualization and modeling of protein, nucleic acid and small molecules. *J Mol Graph* 4, 82–87.
178. Spoel, D. v. d., Lindahl, E., Hess, B., et al. (2005) GROMACS: fast, flexible and free. *J Comp Chem* 26, 1701–1718.
179. Phillips, J. C., Braun, R., Wang, W., et al. (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* 26, 1781–1802.
180. Simons, K. T., Ruczinski, I., Kooperberg, C., et al. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins.* 34, 82–95.
181. Bonneau, R., Tsai, J., Ruczinski, I., et al. (2001) Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins* Suppl. 5, 119–126.
182. Bonneau, R., Strauss, C. E., Rohl, C. A., et al. (2002) De novo prediction of

three-dimensional structures for major protein families. *J Mol Biol* 322, 65–78.

183. Rohl, C. A., Strauss, C. E., Chivian, D., et al. (2004) Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* 55, 656–677.
184. Bradley, P., Malmstrom, L., Qian, B., et al. (2005) Free modeling with Rosetta in CASP6. *Proteins* 61, Suppl. 7, 128–134.
185. Chivian, D., Kim, D. E., Malmstrom, L., et al. (2003) Automated prediction of CASP-5 structures using the Robetta server. *Proteins* 53, Suppl. 6, 524–533.
186. Chivian, D., Kim, D. E., Malmstrom, L., et al. (2005) Prediction of CASP6 structures using automated Robetta protocols. *Proteins* 61, Suppl. 6, 157–166.
187. Kim, D. E., Chivian, D., Baker, D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* 32, W526–531.
188. Vincent, J. J., Tai, C. H., Sathyanarayana, B. K., et al. (2005) Assessment of CASP6 predictions for new and nearly new fold targets. *Proteins* 61, Suppl. 7, 67–83.
189. Wang, G., Jin, Y., Dunbrack, R. L., Jr. (2005) Assessment of fold recognition predictions in CASP6. *Proteins* 61, Suppl. 7, 46–66.
190. Jones, D. T., Bryson, K., Coleman, A., et al. (2005) Prediction of novel and analogous folds using fragment assembly and fold recognition. *Proteins* 61, Suppl. 7, 143–151.
191. Kolinski, A., Bujnicki, J. M. (2005) Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins* 61, Suppl. 7, 84–90.
192. Fujikawa, K., Jin, W., Park, S. J., et al. (2005) Applying a grid technology to protein structure predictor "ROKKY". *Stud Health Technol Inform* 112, 27–36.
193. Debe, D. A., Danzer, J. F., Goddard, W. A., et al. (2006) STRUCTFAST: protein sequence remote homology detection and alignment using novel dynamic programming and profile-profile scoring. *Proteins* 64, 960–967.
194. Ginalski, K., Elofsson, A., Fischer, D., et al. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 19, 1015–1018.
195. Fischer, D. (2003) 3DS3 and 3DS5 3D-SHOTGUN meta-predictors in CAFASP3. *Proteins* 53, Suppl. 6, 517–523.
196. Sasson, I., Fischer, D. (2003) Modeling three-dimensional protein structures for CASP5 using the 3D-SHOTGUN meta-predictors. *Proteins* 53, Suppl. 6, 389–394.
197. Fischer, D. (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins* 51, 434–441.
198. Fischer, D. (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pac Symp Biocomput* 119–130.
199. Lundstrom, J., Rychlewski, L., Bujnicki, J., et al. (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci* 10, 2354–2362.
200. Kurowski, M. A., Bujnicki, J. M. (2003) GeneSilico protein structure prediction metaserver. *Nucleic Acids Res* 31, 3305–3307.
201. Plaxco, K. W., Simons, K. T., Baker, D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 277, 985–994.
202. Bonneau, R., Ruczinski, I., Tsai, J., et al. (2002) Contact order and ab initio protein structure prediction. *Protein Sci* 11, 1937–1944.
203. Shortle, D., Simons, K. T., Baker, D. (1998) Clustering of low-energy conformations near the native structures of small proteins. *Proc Natl Acad Sci U S A* 95, 11158–11162.
204. Venclovas, C., Margelevicius, M. (2005) Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins* 61, Suppl. 7, 99–105.
205. Kosinski, J., Gajda, M. J., Cymerman, I. A., et al. (2005) FRankenstein becomes a cyborg: the automatic recombination and realignment of fold recognition models in CASP6. *Proteins* 61, Suppl. 7, 106–113.
206. Wallner, B., Fang, H., Elofsson, A. (2003) Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. *Proteins* 53, Suppl. 6, 534–541.
207. Wallner, B., Elofsson, A. (2005) Pcons5: combining consensus, structural evaluation and fold recognition scores. *Bioinformatics* 21, 4248–4254.
208. Douguet, D., Labesse, G. (2001) Easier threading through web-based comparisons and cross-validations. *Bioinformatics* 17, 752–753.
209. Takeda-Shitaka, M., Terashi, G., Takaya, D., et al. (2005) Protein structure prediction in CASP6 using CHIMERA and FAMS. *Proteins* 61, Suppl. 7, 122–127.

210. Kopp, J., Schwede, T. (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res* 32, D230–234.
211. Pieper, U., Eswar, N., Braberg, H., et al. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res* 32, D217–222.
212. Yamaguchi, A., Iwadate, M., Suzuki, E., et al. (2003) Enlarged FAMSBASE: protein 3D structure models of genome sequences for 41 species. *Nucleic Acids Res* 31, 463–468.
213. Castrignano, T., De Meo, P. D., Cozzetto, D., et al. (2006) The PMDB Protein Model Database. *Nucleic Acids Res* 34, D306–309.
214. Dayhoff, M. O., Schwartz, R. M., Orcutt, B. C., (1978) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*. M.O. Dayhoff, ed. National Biomedical Research Foundation, Washington, DC.
215. Henikoff, S., Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89, 10915–10919.