Nucleic Acids Research, 2003, Vol. 31, No. 13 ***3375–3380***
DOI: 10.1093/nar/gkg543

Tools for comparative protein structure modeling and analysis

Narayanan Eswar, Bino John[1], Nebojsa Mirkovic[1], Andras Fiser[2], Valentin A. Ilyin[3], Ursula Pieper, Ashley C. Stuart[1], Marc A. Marti-Renom, M. S. Madhusudhan, Bozidar Yerkovich[1] and Andrej Sali*

Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry and California Institute for Quantitative Biomedical Research, University of California, San Francisco, CA USA, [1]Laboratory of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, The Rockefeller University, New York, USA, [2]Department of Biochemistry and Seaver Foundation Center for Bioinformatics, Albert Einstein College of Medicine, New York, USA and [3]Department of Biology, Northeastern University, Boston, MA, USA

Received February 16, 2003; Revised and Accepted March 25, 2003

ABSTRACT

The following resources for comparative protein structure modeling and analysis are described (http://salilab.org): MODELLER, a program for comparative modeling by satisfaction of spatial restraints; MODWEB, a web server for automated comparative modeling that relies on PSI-BLAST, IMPALA and MODELLER; MODLOOP, a web server for automated loop modeling that relies on MODELLER; MOULDER, a CPU intensive protocol of MODWEB for building comparative models based on distant known structures; MODBASE, a comprehensive database of annotated comparative models for all sequences detectably related to a known structure; MODVIEW, a Netscape plugin for Linux that integrates viewing of multiple sequences and structures; and SNPWEB, a web server for structure-based prediction of the functional impact of a single amino acid substitution.

INTRODUCTION

The complete genomes of a number of organisms have been sequenced and many more are under way. Structural biology now faces the arduous task of characterizing the shapes and dynamics of the encoded proteins to facilitate the understanding of their functions and mechanisms of action. Recent developments in the techniques of structure determination at atomic resolution, X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy, have enhanced the quality and speed of structural studies (1). Nevertheless, current statistics still show that the known protein sequences (~1 000 000) (2) vastly outnumber the available protein structures (~20 000) (3). Fortunately, domains in protein sequences are gradually evolving entities that can be clustered into a relatively small number of families of domains with similar sequences and structures (that is, folds) (4). These evolutionary relationships enable the use of computational methods such as threading (5) and comparative protein structure modeling (6,7), to predict the structures of protein sequences based on their similarity to known protein structures. Many structural genomics efforts, in fact, combine experimental structure determination methods and computational modeling techniques to determine enough appropriately selected structures so that most other sequences can be placed within modeling distance of at least one known structure (4,8–10).

Comparative modeling consists of four main steps (7): (i) fold assignment that identifies similarity between the target sequence of interest and at least one known protein structure (the template); (ii) alignment of the target sequence and the template(s); (iii) building a model based on the chosen template(s); and (iv) assessing the model for its accuracy. The accuracy of comparative models is most easily quantified by the extent of sequence similarity between the sequence and the known structure (7,10–12). Accuracy of a model tends to increase with the target-template sequence identity. The errors encountered in comparative modeling include fold assignment and alignment errors (which occur mostly below 30% sequence identity), distortions and shifts in the core segments and loops, as well as errors in side-chain packing (which occur in varying degrees throughout the spectrum of sequence similarity).

A number of servers for automated comparative modeling are available (http://salilab.org/bioinformatics_resources.shtml). Automation makes comparative modeling accessible to both experts and nonspecialists alike. Many of the servers are tested at the biannual CAFASP meetings (13) and continually by the LiveBench (14) and EVA (15,16) web servers for assessment of automated structure prediction methods. However, in spite of automation, the process of calculating a model for a given

*To whom correspondence should be addressed at: Mission Bay Genentech Hall, Suite N472D, 600 16th Street, University of California, San Francisco, CA 94143-2240, USA. Tel: +1 4155144227; Fax: +1 4155144231; Email: sali@salilab.org

 © *Oxford University Press 2003; all rights reserved*

sequence, refining its accuracy as well as visualizing and analysing its family members in sequence and structure space can involve the use of scripts, local programs and servers scattered across the internet and not necessarily interconnected. In addition, manual intervention is generally still needed to maximize the accuracy of the models in the difficult cases. We present here our resources, available through http://salilab.org, that begin to address these shortcomings.

SOFTWARE AND WEB SERVERS

MODELLER

MODELLER is a computer program for comparative protein structure modeling (http://salilab.org/modeller) (17,18). In the simplest case, the input is an alignment of a sequence to be modeled with the template structures, the atomic coordinates of the templates and a short script file. MODELLER then automatically calculates a model containing all non-hydrogen atoms, without any user intervention and within minutes on a Pentium processor.

MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints (17). The spatial restraints include: (i) homology-derived restraints on the distances and dihedral angles in the target sequence, extracted from its alignment with the template structures (17); (ii) stereochemical restraints such as bond length and bond angle preferences, obtained from the CHARMM-22 molecular mechanics forcefield (19); (iii) statistical preferences for dihedral angles and non-bonded interatomic distances, obtained from a representative set of known protein structures (20); and (iv) optional manually curated restraints, such as those from NMR spectroscopy, rules of secondary structure packing, cross-linking experiments, fluorescence spectroscopy, image reconstruction from electron microscopy, site-directed mutagenesis and intuition. The spatial restraints, expressed as probability density functions, are combined into an objective function that is optimized by a combination of conjugate gradients and molecular dynamics with simulated annealing. This model building procedure is similar to structure determination by NMR spectroscopy.

Apart from model building, MODELLER can perform additional auxiliary tasks, including alignment of two protein sequences or their profiles, multiple alignment of protein sequences and/or structures, calculation of phylogenetic trees and *de novo* modeling of loops in protein structures (18). MODELLER is written in Fortran 90 and runs on Pentium PCs (Linux and Windows XP), Apple Macintosh (OS X) and workstations from Silicon Graphics (IRIX), Sun (Solaris), IBM (AIX) and DEC Alpha (OSF/1). The program is used with its own scripting language and does not include a graphical interface.

MODWEB

MODWEB is a web server for automated comparative protein structure modeling (http://salilab.org/modweb) (12). MODWEB accepts one or many sequences in the FASTA format (21) and calculates models for them based on the best available template structures from the Protein Data Bank (PDB) (http://www.rcsb.org) (3). Alternatively, MODWEB also accepts a protein structure as an input and calculates models for all its identifiable sequence homologs in the non-redundant SWISS-PROT protein sequence database (2). The latter mode is a useful tool for various structural genomics efforts to assess the impact of a newly determined structure on the modeling coverage of the sequence space (4,8). The New York Structural Genomics Research Consortium routinely processes all newly determined structures using this approach and the results are available through MODBASE (22) and at http://nysgxrc.org/nysgrc/mod_results.html.

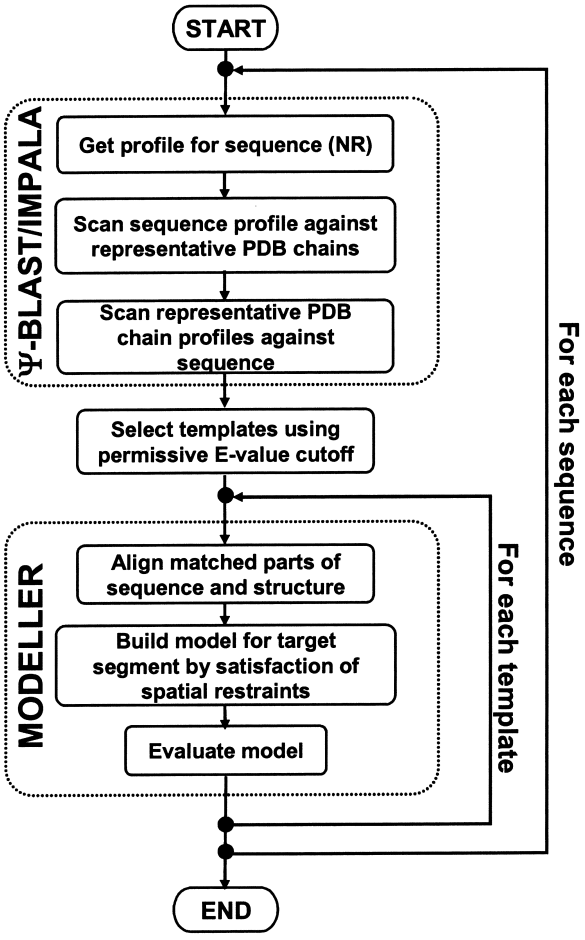

Figure 1. Flowchart of MODPIPE, a large-scale protein structure modeling pipeline. See text for details.

MODWEB relies on MODPIPE, a completely automated software pipeline for comparative protein structure modeling, which can calculate comparative models for a large number of protein sequences, using many different template structures and sequence–structure alignments (7,12,22) (Fig. 1). Sequence–structure matches are established by aligning the PSI-BLAST sequence profile (23) of the target sequence against each of the template sequences extracted from PDB (3), as well as by scanning the target sequence against a database of the template profiles using IMPALA (24). Significant alignments covering distinct regions of the target sequence are chosen for modeling. Models are calculated for each of the sequence–structure matches

using MODELLER (17). The resulting models are then evaluated by a composite model quality criterion (below) (25).

The thoroughness of a search for the best model is modulated by a number of user parameters, including two E-value thresholds for identifying useful sequence–structure relationships and the degree of conformational sampling given a sequence–structure alignment. The validity of sequence–structure relationships is not prejudged at the fold detection stage, but is assessed after the construction of the model and its evaluation. This approach enables a thorough exploration of fold assignments, sequence–structure alignments and conformations, with the aim of finding the model with the best evaluation score.

For single sequences, MODWEB returns the output of the calculations by email, but when the input consists of many sequences or a structure, the output is added as a separate dataset into our relational database of protein structure models, MODBASE (22), described below.

MODLOOP

MODLOOP is a web server for explicit modeling of loops in protein structures (http://salilab.org/modloop) (18) (A. Fiser and A. Sali, submitted for publication). The server requires an input consisting of a coordinate file (PDB format) and the starting and ending residue positions of the loops. The user can also specify several loops to be optimized simultaneously, which is particularly useful in the case of multiple interacting loops. The prediction is returned to the user by email.

MODLOOP is a front-end to prediction of loop conformations by MODELLER (18). Spatial restraints are obtained as described for MODELLER, except for the absence of the homology-derived restraints. The environment of the loop(s) is fixed and does not change during optimization. The protocol generates 300 loop predictions by starting with random initial loop conformations, followed by a thorough optimization of the positions of all non-hydrogen atoms. The conformation with the lowest objective function score is chosen as the final loop prediction. Depending on the accuracy of the loop environment, predictions for single loops as long as 12 residues may be of useful accuracy (18).

To make the server respond more rapidly, MODLOOP submits individual calculations to a cluster of PCs running Linux. In addition, to limit the load on our computational resources, the number of independently calculated loop conformations is restricted to 300 and the total length of all specified loops cannot be more than 20 residues.

MOULDER

MOULDER is an optional protocol available to the MODWEB web server (http://salilab.org/modweb) (B. John and A. Sali, submitted for publication). If chosen, an iteration of target–template alignment, model building and model assessment replaces the default model building step by MODELLER in the standard MODPIPE protocol. Other input and output specifications are as described for MODWEB.

MOULDER optimizes both the given alignment and the model implied by it. The optimization relies on a genetic algorithm protocol that starts with an initial alignment and then iterates through realignment, model building and model assessment to optimize a model assessment score. During this iterative process (i) new alignments are constructed by application of genetic algorithm operators, such as alignment mutation and cross-over; (ii) the comparative models corresponding to these alignments are built by satisfaction of spatial restraints, as implemented in MODELLER; and (iii) the models are assessed by a composite criterion, partly depending on an atomic statistical potential. This iterative approach blurs the boundary between traditional comparative modeling, which calculates a highly refined model for one alignment, and threading, which calculates a simple implicit model for each one of the many tested alignments.

MOULDER runs on a cluster of computers running the Linux operating system. For a 150-residue target sequence, the protocol currently requires approximately a day of computation on 100 CPUs. Because of this demanding computational load, the MOULDER option of MODWEB is currently restricted to a small number of selected users.

MODBASE

MODBASE is a comprehensive relational database of annotated comparative protein structure models. It contains several model datasets, including that for all available protein sequences matched to at least one known protein structure (http://salilab.org/modbase) (22,26,27). This dataset was calculated by applying MODPIPE to all sequences in the SWISS-PROT database (March 2002) (2). Currently, MODBASE contains models for domains in 415 937 out of 733 239 (~57%) unique protein sequences found in SWISS-PROT.

MODBASE is queryable through its web user interface by PDB codes, SWISS-PROT and GENPEPT accession numbers, open reading frame names, various keywords, model reliability, model size, target–template sequence identity, alignment significance and sequence similarity against the modeled sequences as detected by BLAST (23). It is also possible to query the database directly using SQL as implemented in MySQL.

The output of a search is displayed on pages with varying amounts of information about the modeled sequences, template structures, alignments, and functional annotations. These tables also contain links to other sequence, structure and function annotation databases, such as PDB (3), GenBank (28), SWISS-PROT (2), CATH (29), PFAM (30) and PRODOM (31). In addition to the web pages containing text and schematic representations implemented in PERL/CGI, MODBASE uses the Netscape plugin MODVIEW (32), described below, to visualize and analyse the models of target sequences, template structures and their alignments.

ACCURACY OF COMPARATIVE MODELS

The accuracy of comparative models calculated by MODWEB, MOULDER and those in MODBASE can be judged by a variety of criteria (7,10), including percentage sequence identity on which the models are based (12) and various model assessment scores (25,33). In particular, we use a composite GA341 model score that combines a Z-score calculated with a statistical potential function (25), target–

template sequence identity and a measure of structural compactness (Melo,F. *et al.*, in preparation). The GA341 score ranges from 0 for models that tend to have an incorrect fold to 1 for models that tend to be comparable to low-resolution X-ray structures. Comparison of models with their corresponding experimental structures indicates that models with GA341 scores greater than 0.7 generally have the correct fold with more than 35% of the backbone atoms superposable within 3.5 Å. Reliable models (GA341 score $\geq$0.7) based on alignments with more than 40% sequence identity have a median overlap of more than 90% with the corresponding experimental structure. In the 30–40% sequence identity range, the overlap is usually between 75 and 90% and below 30% it drops to 50–75%, or even less in the worst cases.

MODVIEW

MODVIEW is a Netscape plug-in for Linux that integrates a multiple structure viewer, a multiple sequence alignment editor and a database querying engine (http://salilab.org/modview) (32). A user can interactively manipulate hundreds of proteins, visualize conserved and variable residues, active and binding sites, fragments and domains in protein families, as well as display large macromolecular complexes such as ribosomes or viruses. As a Netscape plug-in, MODVIEW can be included in HTML pages along with text and figures, which makes it useful for teaching and presentations. MODVIEW is also suitable as a graphical interface to various databases because it can be controlled through JavaScript commands and called from CGI scripts.

Structure visualization in MODVIEW is based on the program MIPA (34), with graphical rendering and scripting from RasMol (35). The sequence and alignment editor is based on JalView (http://www2.ebi.ac.uk/~michele/jalview). MODVIEW has a number of options that control the display of structures, selection of atoms, presentation styles and coloring schemes. It is possible to superpose a set of structures with a single click, either by structure or according to a given alignment. Each structure in the display or a subset of atoms can be manipulated independently from the remaining structures. MODVIEW can also be used to create pairwise and multiple sequence alignments. The alignments can be edited manually, and the sequences can be compared with the aid of dendrograms and principal component analysis.

SNPWEB

SNPWEB is a web server for prediction of the functional effect of a single amino acid residue substitution (http://salilab.org/SNPWeb) (N. Mirkovic, M.A. Marti-Renom, A. Sali and A. Monteiro, submitted for publication). The server takes as input the specifications of the wild-type protein structure and a single amino acid residue substitution. The output, in a matter of minutes, is a prediction of whether or not the function of the mutant is impaired, as well as the rationalization of the predicted impact in terms of several features of the wild-type and mutant structures.

The specified wild-type structure is first loaded or located in PDB or MODBASE; if not found, modeling is attempted with MODWEB. If the wild-type structure cannot be obtained in any one of these four ways, the calculation cannot proceed and no results are displayed. Alternatively, with the wild-type structure in hand, the model of the mutant is calculated by the MUTATE_MODEL command of MODELLER. Next, the server calculates a set of sequence- and structure-based features for the wild-type and mutant proteins, including: (i) accessible surface area (ASA); (ii) rigidity of the changed position and its neighborhood as indicated by the average isotropic temperature factor from X-ray crystallography of the wild-type structure or the template structure used to calculate the wild-type model; (iii) changes in residue volume, ASA, charge and hydrophobicity; (iv) the degree of evolutionary conservation at the replacement position among the members of the corresponding sequence family; and (v) the replacement likelihood from the family-specific substitution matrix (36). Optionally, additional features, such as the location of the substitution relative to known functional site(s) and known structural and/or functional importance of the residues, may also be included in the feature set for the special cases curated by hand. The subsequent classification of the mutation as neutral or deleterious is achieved by a decision tree. The protocol is based on the assumption that a mutation is deleterious in either one of the following two ways: (i) when it is exposed to the solvent, it may substantially change the structure or chemical nature of functional sites that bind other molecules; or (ii) when it is buried in the core, it may prevent folding of the domains into their native fold, or, less likely, affect only the structure of functional sites. In the case of the human BRCA1 domains, the server is able to rationalize 31 of 37 point mutations with known functional impact.

DISCUSSION

The process of comparative protein structure modeling usually requires the use of many programs, to identify template structures, to generate sequence–structure alignments, to build the models and to evaluate them. In addition, various sequence and structure databases that are accessed by these programs are needed. Once an initial model is calculated, it is generally refined and finally analysed in the context of many other related proteins and their functional annotations. To facilitate these tasks, we developed several programs, servers and databases (Table 1), some of which have been described above. These resources can be divided into four main categories.

Firstly, initial construction of comparative models is achieved by MODELLER, MODPIPE and MODWEB. MODELLER is the computational engine for our other modeling programs and servers. MODPIPE automates the entire comparative modeling process, relying on PSI-BLAST, IMPALA and MODELLER for its functionality. MODWEB provides a simple web interface to MODPIPE to make it accessible to any user.

Secondly, refinement of comparative models is facilitated by MODELLER, MOULDER and MODLOOP. An expert may utilize the MOULDER protocol in MODWEB to attempt to overcome errors in sequence–structure alignments between distantly related proteins. The user may also use the MODLOOP server to refine the conformation of loops, such as those calculated by MODWEB.

Table 1. List of our computer programs, Web servers and databases

Name	Type	URL, references	Input	Output/contents
MODELLER	Program	http://salilab.org/modeller (17,18)	• Sequence–structure alignment • Structure files • Script file	Comparative model of the target sequence, with all nonhydrogen atoms
MODPIPE	Program	(12)	• One or many sequences	Comparative models for the sequences
MODWEB	Web server	http://salilab.org/modweb (12)	• One or many sequences, or • Structure file	Comparative models for the sequences or sequence homologs of the input structure
MODLOOP	Web server	http://salilab.org/modloop (18)	• Structure file • Starting and end points of the loops to be modeled	Coordinates of the input structure with optimized loop conformation
MOULDER	Protocol	http://salilab.org/modweb (B.John and A.Sali, submitted for publication)	• See MODWEB	See MODWEB
SNPWEB	Web server	http://salilab.org/SNPWeb (Mirkovic *et al.*, submitted for publication)	• Native structure specification • Substitution specification	Functional impact of the residue substitution
EVA	Web server	http://salilab.org/eva (16)	• Coordinate files of structure predictions from automated servers	Evaluation of the servers and ranking of their accuracies
MODBASE	Database	http://salilab.org/modbase (22)		Fold assignments, sequence–structure alignments, models, model assessments for all sequences related to known structures
LIGBASE	Database	http://salilab.org/ligbase (37)		All families of related binding sites of known structure
DBAli	Database	http://salilab.org/DBAli (38)		Many multiple structure-based alignments
ICEDB	Database, LIMS	http://salilab.org/icedb		Target tracking system for structural genomics

LIMS: laboratory information management system. The common formats for the sequence and structure files are those of FASTA and PDB, respectively.

Thirdly, inspection and analysis of models is made easier by MODBASE, MODVIEW, SNPWEB and LIGBASE (37). Comparative modeling provides only a starting point for annotating the function of a protein, using other theoretical and experimental techniques. MODBASE stores comparative models for all protein sequences detectably related to a known protein structure and additionally links each model to a number of external databases. MODVIEW can be used to visualize the models and alignments stored in the databases. LIGBASE is a database comprising all ligand-binding sites of known structure aligned with all related protein sequences and structures, and may be useful in the examination of putative binding sites on the models in MODBASE. SNPWEB relies on protein structure considerations to predict the functional impact of a given amino acid mutation.

Fourthly, improvement of the modeling techniques is served by DBALI, a comprehensive database of multiple protein structure alignments calculated by the SALIGN command of MODELLER (38) and the EVA-CM web server for automated and continuous assessment of comparative protein structure modeling web servers (15,16).

These resources, in combination with those from others, will contribute to structure-based functional annotation of proteins and thus enhance the impact of genome sequencing, structural genomics and functional genomics on biology and medicine.

ACKNOWLEDGEMENTS

This research was supported by the grants NIH/NIGMS R01 GM 54762, NIH/NIGMS P50 GM62529, NIH/NCI R33 CA84699, Merck Genome Research Institute Award, Mathers Foundation Award and Sun Academic Equipment Grant EDUD-7824-020257-US. A.S. is an Irma T. Hirschl Trust Career Scientist.

REFERENCES

1. Zhang,C. and Kim,S.H. (2003) Overview of structural genomics: from structure to function. *Curr. Opin. Chem. Biol.*, **7**, 28–32.
2. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
3. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D. Biol. Crystallogr.*, **58**, 899–907.

4. Vitkup,D., Melamud,E., Moult,J. and Sander,C. (2001) Completeness in structural genomics. *Nature Struct. Biol.*, **8**, 559–566.
5. Domingues,F.S., Lackner,P., Andreeva,A. and Sippl,M.J. (2000) Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol.*, **297**, 1003–1013.
6. Blundell,T.L., Sibanda,B.L., Sternberg,M.J. and Thornton,J.M. (1987) Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, **326**, 347–352.
7. Marti-Renom,M.A., Stuart,A.C., Fiser,A., Sanchez,R., Melo,F. and Sali,A. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.
8. Sali,A. (1998) 100,000 protein structures for the biologist. *Nature Struct. Biol.*, **5**, 1029–1032.
9. Sanchez,R., Pieper,U., Melo,F., Eswar,N., Marti-Renom,M.A., Madhusudhan,M.S., Mirkovic,N. and Sali,A. (2000) Protein structure modeling for structural genomics. *Nature Struct. Biol.*, **7** (suppl.), 986–990.
10. Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
11. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
12. Sanchez,R. and Sali,A. (1998) Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc. Natl Acad. Sci. USA*, **95**, 13597–13602.
13. Fischer,D., Elofsson,A., Rychlewski,L., Pazos,F., Valencia,A., Rost,B., Ortiz,A.R. and Dunbrack,R.L., Jr. (2001) CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins*, **5** (suppl.), 171–183.
14. Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) LiveBench-2: large-scale automated evaluation of protein structure prediction servers. *Proteins*, **5** (suppl.), 184–191.
15. Koh,I.Y.Y., Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Eswar,N., Graña,O., Pazos,F., Valencia,A., Sali,A. and Rost,B. (2003) EVA: evaluation of protein structure prediction servers. *Nucleic Acids Res.*, **31**, 3311–3315.
16. Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Fiser,A., Pazos,F., Valencia,A., Sali,A. and Rost,B. (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, **17**, 1242–1243.
17. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
18. Fiser,A., Do,R.K. and Sali,A. (2000) Modeling of loops in protein structures. *Protein Sci.*, **9**, 1753–1773.
19. MacKerell,A.D.J., Bashford,D., Bellott,R.L., Dunbrack,R.L., Jr., Evanseck,J.D., Field,M.J., Fischer,S., Gao,J., Guo,H., Ha,S. *et al.* (1998) All-Atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.*, **102**, 3586–3616.
20. Sali,A. and Overington,J.P. (1994) Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.*, **3**, 1582–1596.
21. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.
22. Pieper,U., Eswar,N., Stuart,A.C., Ilyin,V.A. and Sali,A. (2002) MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.*, **30**, 255–259.
23. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
24. Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
25. Melo,F., Sanchez,R. and Sali,A. (2002) Statistical potentials for fold assessment. *Protein Sci.*, **11**, 430–448.
26. Sanchez,R. and Sali,A. (1999) ModBase: a database of comparative protein structure models. *Bioinformatics*, **15**, 1060–1061.
27. Sanchez,R., Pieper,U., Mirkovic,N., de Bakker,P.I., Wittenstein,E. and Sali,A. (2000) MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.*, **28**, 250–253.
28. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
29. Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. and Orengo,C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.
30. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
31. Servant,F., Bru,C., Carrere,S., Courcelle,E., Gouzy,J., Peyruc,D. and Kahn,D. (2002) ProDom: automated clustering of homologous domains. *Brief. Bioinform.*, **3**, 246–251.
32. Ilyin,V.A., Pieper,U., Stuart,A.C., Marti-Renom,M.A., McMahan,L. and Sali,A. (2003) ModView, visualization of multiple protein sequences and structures. *Bioinformatics*, **19**, 165–166.
33. Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.
34. Ilyin,V.A. (1994) Non-polar nuclei in fungal microbial RNases. *Protein Eng.*, **7**, 1189–1195.
35. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.
36. Henikoff,J.G. and Henikoff,S. (1996) Using substitution probabilities to improve position-specific scoring matrices. *Comput. Appl. Biosci.*, **12**, 135–143.
37. Stuart,A.C., Ilyin,V.A. and Sali,A. (2002) LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics*, **18**, 200–201.
38. Marti-Renom,M.A., Ilyin,V.A. and Sali,A. (2001) DBAli: a database of protein structure alignments. *Bioinformatics*, **17**, 746–747.