Nucleic Acids Research, 2002, Vol. 30, No. 1 255–259

MODBASE, a database of annotated comparative protein structure models

Ursula Pieper, Narayanan Eswar, Ashley C. Stuart, Valentin A. Ilyin and Andrej Sali*

Laboratories of Molecular Biophysics, The Pels Family Center for Biochemistry and Structural Biology, The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA

Received September 18, 2001; Revised and Accepted October 2, 2001

ABSTRACT

MODBASE (http://guitar.rockefeller.edu/modbase) is a relational database of annotated comparative protein structure models for all available protein sequences matched to at least one known protein structure. The models are calculated by MODPIPE, an automated modeling pipeline that relies on PSI-BLAST, IMPALA and MODELLER. MODBASE uses the MySQL relational database management system for flexible and efficient querying, and the MODVIEW Netscape plugin for viewing and manipulating multiple sequences and structures. It is updated regularly to reflect the growth of the protein sequence and structure databases, as well as improvements in the software for calculating the models. For ease of access, MODBASE is organized into different datasets. The largest dataset contains models for domains in 304 517 out of 539 171 unique protein sequences in the complete TrEMBL database (23 March 2001); only models based on significant alignments (PSI-BLAST E-value < 10^{-4}) and models assessed to have the correct fold are included. Other datasets include models for target selection and structure-based annotation by the New York Structural Genomics Research Consortium, models for prediction of genes in the *Drosophila melanogaster* genome, models for structure determination of several ribosomal particles and models calculated by the MODWEB comparative modeling web server.

INTRODUCTION

The genome sequencing projects are providing us with complete sets of amino acid sequences of many proteins, including catalysts, inhibitors, messengers, receptors, transporters and structural elements of living organisms. To realize the full potential of the genome projects, we need to be able to assign, understand, control and modify the function of the proteins encoded by the genomes. These tasks are generally facilitated by the knowledge of the native three-dimensional (3D) structure of the proteins (1–3). Unfortunately, the structures of only a tiny fraction of known protein sequences have been defined by X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR). There are only about 16 000 entries in the Protein Data Bank (PDB) of known protein structures (4), whereas there are over 600 000 entries in the comprehensive TrEMBL (5) and GenPept (6) protein sequence databases. Therefore, the structure of most protein sequences has to be predicted by computation (7). There are two classes of protein structure prediction methods. The first class of methods, *de novo* or *ab initio* methods, predicts the structure from sequence alone, without relying on similarity at the fold level between the modeled sequence and any of the known structures (8). Despite significant recent progress, *de novo* prediction is not yet generally applicable because even successful calculations result in models with a root-mean-square (r.m.s.) error of 4–8 Å over approximately 80 residues. The second class of protein structure prediction methods, including threading and comparative modeling, rely on detectable similarity spanning most of the modeled sequence and at least one known structure (9). Comparative modeling consists of four steps: finding known structures related to the sequence to be modeled (i.e. templates), aligning the sequence with the templates, building a model and assessing the model. Next, we describe the errors in comparative models, applications of comparative models and the current coverage of genomes by comparative models.

The accuracy of comparative modeling is related to the percentage sequence identity on which the model is based, correlating with the relationship between the structural and sequence similarities of two proteins (9–11). High accuracy comparative models are based on >50% sequence identity to their templates. They tend to have ~1 Å r.m.s. error for the main-chain atoms, which is comparable to the accuracy of a medium-resolution NMR structure or a low-resolution X-ray structure. The errors are mostly mistakes in side-chain packing, small shifts or distortions of the core main-chain regions and occasionally larger errors in loops. Medium-accuracy comparative models are based on 30–50% sequence identity. They tend to have ~90% of the main-chain modeled with 1.5 Å r.m.s. error. There are more frequent side-chain packing, core distortion and loop modeling errors, and there are occasional alignment mistakes. And finally, low accuracy comparative models are based on <30% sequence identity. The alignment errors increase rapidly below 30% sequence identity and become the most significant origin of errors in comparative models. In addition, when a model is based on an almost insignificant alignment to a known structure, it may also have an entirely incorrect fold. Errors in comparative modeling and

*To whom correspondence should be addressed. Tel: +1 212 327 7550; Fax: +1 212 327 7540; Email: sali@rockefeller.edu

Table 1. Summary of some datasets of models in ModBase

Dataset	Number of sequences attempted to be modeled	Number of sequences with reliable fold assignments	Number of models	Access
TrEMBL (2000)	415 801	197 999	371 816	Academic
TrEMBL (2001)	539 171	304 517	625 739	Academic
Homo sapiens	33 093	19 437	53 965	
Mus musculus	20 792	11 772	32 138	
Drosophila melanogaster	16 567	8692	27 240	
Caenorhabditis elegans	19 326	9538	26 083	
Arabidopsis thaliana	29 213	16 052	41 164	
Saccharomyces cerevisiae	6714	2972	7218	
Escherichia coli	13 787	6336	11 572	
Mycoplasma genitalium	564	285	533	
MODWEB	~4500	3994	5140	Private
NYSGRC	13 451	7956	27 886	NYSGRC
*Drosophila melanogaster**	21 225	7112	200 153	Academic
Saccharomyces cerevisiae ribosome	109	80	221	Academic

The number of sequences attempted to be modeled indicates the number of original sequences submitted to MODPIPE. For a definition of a reliable fold assignment see 'Contents'. The number of models can be larger than the number of sequences because different segments of a sequence may be modeled independently and because the same segment may be modeled based on different template structures. The two TrEMBL datasets correspond to the June 2000 and March 23, 2001 versions of the complete TrEMBL database, respectively. For the 2001 TrEMBL dataset, the numbers for several organisms are shown separately. These numbers correspond to all the entries in the TrEMBL database, including multiple submissions, mutants and partial sequences. The MODWEB datasets are created by the MODWEB comparative modeling web server (http://guitar.rockefeller.edu/modweb) (N.Eswar and A.Sali, manuscript in preparation). The NYSGRC datasets are used in target selection and structure-based annotation by NYSGRC (37). The *D.melanogaster** dataset contains models for the over-predicted putative genes in the *D.melanogaster* genome (38). The *S.cerevisiae* ribosome dataset contains comparative models for proteins in the yeast ribosome (40).

threading are best quantified by continuous, automated and large-scale assessment of automated prediction methods, such as the assessment by the LiveBench (12) and EVA web servers (13). Accuracies of the best model building methods are relatively similar when used optimally (11,14). Other factors such as template selection and alignment accuracy usually have a larger impact on the model accuracy, especially for models based on <40% sequence identity to the templates.

Reasonable applications of any protein structure model depend on its accuracy, and even models with large errors can be helpful (7,15). For example, high- and medium-accuracy comparative models are frequently useful in refining functional predictions that are based on a sequence match alone, because ligand binding is more directly determined by the structure of the binding site than by its sequence. It is often possible to correctly predict features of the target protein that do not occur in the template structure. For example, the size of a ligand may be predicted from the volume of the binding site cleft (16) and the location of a binding site for a charged ligand can be predicted from a cluster of charged residues on the protein (17). Fortunately, errors in the functionally important regions in comparative models tend to be relatively low because the functional regions, such as active sites, tend to be more conserved in evolution than the rest of the fold (18). Comparative models have been used in studying catalytic mechanisms of enzymes, designing and improving ligands, docking of macromolecules, prediction of interacting protein partners, virtual screening and docking of small ligands, defining antibody epitopes, molecular replacement in X-ray crystallography, designing chimeras, stable and crystallizable variants, supporting site-directed mutagenesis, refining NMR structures, fitting proteins into low-resolution electron density maps, finding functional sites by 3D motif searching, determining structure from sparse experimental restraints, annotating function by fold assignment and establishing evolutionary relationships.

While the models can provide substantial insight, they can also be misleading. Thus, it is critical that each model be assessed prior to its use. In general, the accuracy of a comparative model can be estimated simply from sequence similarity to its template (9,10), or more generally by a variety of model assessment methods (9,19–21).

Threading and comparative modeling methods have already been applied on a genomic scale (10,22–27). The fraction of the known protein sequences that have at least one segment detectably related to one or more known structures varies with a genome, and currently ranges from 20 to 65% (28). Approximately 57% of all non-redundant protein sequences in the TrEMBL database have at least one domain that can be characterized structurally by comparative modeling (Table 1). Thus, the number of sequences that can be modeled with useful accuracy by comparative modeling is already more than an order of magnitude larger than the number of experimentally determined protein structures. While the current number of modeled proteins may look impressive, usually only one domain per protein is modeled (on average, proteins have slightly more than two domains) and two-thirds of the models are based on <30% sequence identity to the closest template. However, the accuracy and applicability of comparative modeling are improving rapidly, primarily reflecting the growth of the number and variety of the known protein structures, determined both by

small teams of structural biologists as well as the world-wide effort in structural genomics (29–31).

Comparative modeling is already a significant method in biology because a large fraction of proteins can be modeled with accuracy that is sufficient for addressing many biological questions. To increase the efficiency of using comparative models for experts and to make comparative models accessible to non-experts, we developed MODBASE, a comprehensive database of comparative models for all protein sequences that are detectably related to proteins of known structure (32,33). In this paper, we describe the most recent version of MODBASE.

CONTENTS

Comparative models in MODBASE are calculated using MODPIPE, the entirely automated software pipeline for large-scale comparative protein structure modeling (10; N.Eswar, R.Sanchez, M.A.Marti-Renom, M.S.Madhusudhan, F.Melo, U.Pieper, A.C.Stuart, V.A.Ilyin and A.Sali, manuscript in preparation). MODPIPE relies on PSI-BLAST (34) and IMPALA (35) for fold assignment, and the MODELLER package for sequence–structure alignment, model building and model assessment (36). MODBASE currently contains fold assignments, sequence–structure alignments, all-atom comparative models, and model assessments for segments of approximately 350 000 protein sequences. Fold assignments and models for a fraction of these sequences are considered unreliable. The folds of the models are assessed by computing an energy-based model score that uses a statistical energy function, sequence similarity with the modeling template and a measure of structural compactness (9,21). Tests with known structures have shown that models with scores from 0.7 to 1.0 have the correct fold at a 95% confidence level.

For ease of access, the contents of MODBASE is organized into several datasets (Table 1). The largest of the datasets was obtained by processing all of the 539 171 unique protein sequences in the SWISS-PROT, TrEMBL and TrEMBL-NEW databases (March 23, 2001). The entire calculation took ~6 weeks of CPU time on a Linux cluster with 340 Pentium III CPUs. The TrEMBL dataset contains only reliable fold assignments, corresponding to either reliable models or models based on a reliable PSI-BLAST match. A model is reliable when its energy-based model score is >0.7. A PSI-BLAST match is reliable when the corresponding E-value from a filtered PSI-BLAST search is $<10^{-4}$.

A large number of MODBASE datasets are created by the web server for automated comparative protein structure modeling, MODWEB (http://guitar.rockefeller.edu/modweb) (N.Eswar and A.Sali, manuscript in preparation). MODWEB provides a web interface to MODPIPE and takes as input either a set of sequences or a protein structure. For all input sequences, models are calculated when a potentially related known protein structure is found in PDB. For an input protein structure, models are produced for all the detectably related protein sequences in a comprehensive non-redundant sequence database. MODBASE provides convenient storage and access to the models calculated by MODWEB.

MODBASE is also used in target selection and structure-based annotation by the New York Structural Genomics Research Consortium (NYSGRC). For target selection, MODBASE provides information about protein sequences that cannot be satisfactorily modeled by comparative modeling. For structure-based annotation, MODBASE contains models calculated by MODWEB for all sequences detectably related to the novel X-ray structures from the NYSGRC (37). In addition, the NYSGRC measures the impact of its structures by documenting the number and quality of the corresponding models for detectably related proteins in the non-redundant sequence database. For each new structure, an average of approximately 100 protein sequences without any prior structural characterization are modeled at least at the fold level (http://www.nysgrc.org/). This large leverage of structure determination by protein structure modeling illustrates and justifies the premise of structural genomics.

Another application of MODBASE was to facilitate prediction of genes in the *Drosophila melanogaster* genome (38). In the first step, twice the expected number of genes were predicted by using GeneScan with promiscuous parameters (39). In the second step, each of the putative genes was tested by a variety of criteria to detect potentially valid gene predictions. One of these criteria was protein sequence similarity between a putative gene and any of the known protein sequences. This information was calculated for all of the putative genes by MODPIPE and was stored in MODBASE.

Yet another application of MODBASE was to facilitate construction of a molecular model of the yeast ribosomal particle (40). The molecular model of the whole yeast ribosome was calculated by fitting protein models extracted from MODBASE into the electron density of the 80S ribosomal particle, obtained by electron microscopy at 15 Å resolution. Most of the models for 40 out of the 75 ribosomal proteins were based on ~30% sequence identity to their template structures. This example also suggests that structural genomics of single proteins or their domains, combined with protein structure prediction, may contribute significantly to efficient structural characterization of large macromolecular assemblies.

An example of how MODBASE can be used to elucidate function of a specific protein is provided by the identification and characterization of a *p53* homolog in *D.melanogaster* (*dp53*) (41). A simple query of MODBASE revealed an uncharacterized *D.melanogaster* protein with segments that could be modeled reliably based on the known structure of the human *p53* protein, despite low sequence similarity between the two proteins. An inspection of the corresponding alignment and comparative model showed that many of the residues known to be important for the function of the human protein were conserved in the putative *dp53* (i.e. the DNA binding residues). This observation justified extensive characterization of the biochemical and cellular functions of the *dp53*, both *in vitro* and *in vivo*, proving that the roles of the human and *D.melanogaster p53* proteins are indeed similar. Hence, *D.melanogaster* may provide a useful, simpler genetic system to further study the *p53* regulation network.

ACCESS AND INTERFACE

MODBASE is queryable through the web at http://guitar.rockefeller.edu/modbase by PDB codes, SWISS-PROT/TrEMBL 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 (34). It is also possible to

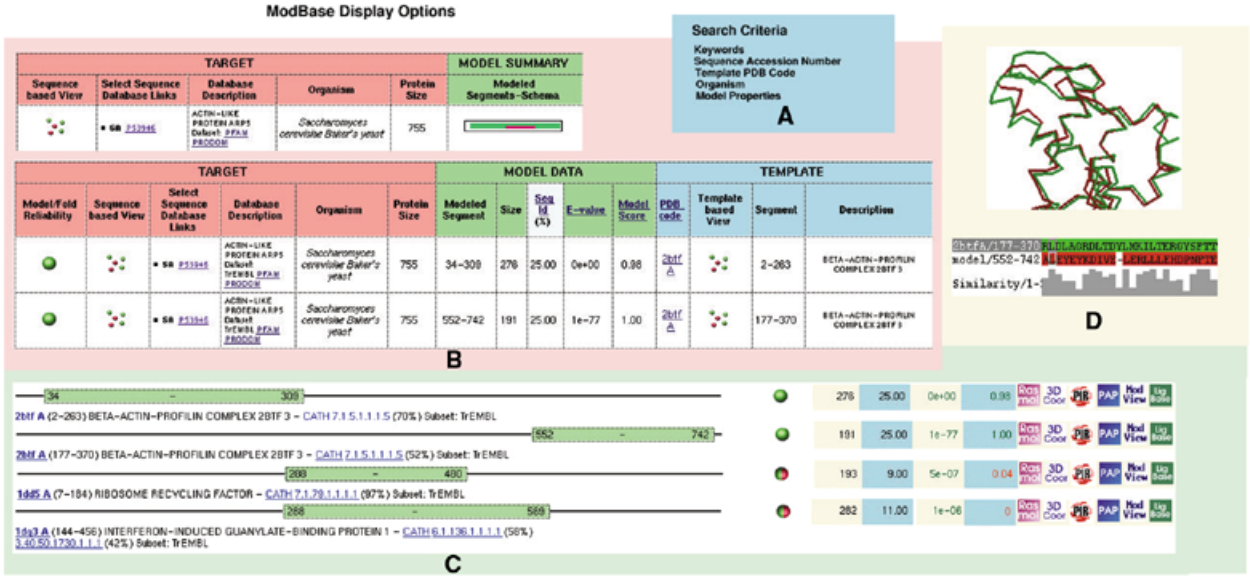

Figure 1. Some MODBASE query and results pages. (**A**) Summary of search criteria. (**B**) Summary display of search results, either all models satisfying the search criteria or all sequences with models satisfying the search criteria. (**C**) Model display page shows a schematic alignment and some information either about all models of one sequence or about all models based on one template structure. (**D**) Sample window of MODVIEW, a Netscape plugin for displaying and analyzing multiple sequences and structures (http://guitar.rockefeller.edu/modview).

query the database directly using SQL as implemented in MySQL (http://www.mysql.com). While access to MODBASE is free for academic researchers, it is regulated by a login procedure that relies on cookies and restricts access to certain datasets. For example, models calculated by a MODWEB user are not accessible to others, and preliminary datasets, such as the models produced with unreleased NYSGRC structures, are also protected.

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 (4), GenBank (6), TrEMBL (5), CATH (42), Pfam (43) and ProDom (44). In addition, MODBASE is linked to LIGBASE (http://guitar.rockefeller.edu/ligbase) (45), our database comprising ligand-binding sites of known structure aligned with related protein sequences and structures. Currently, LIGBASE contains approximately 50 000 ligand binding sites for small molecules found in the PDB. Binding sites are defined by protein atoms within 5 Å of any ligand atom. The link between MODBASE and LIGBASE allows display of putative ligand binding residues in those MODBASE models that can be related to the protein structures with defined binding sites, as established by the structural alignments from the CE program (46) and sequence–structure alignments in MODBASE.

In addition to the web pages containing text and schematic representations implemented in Perl/CGI, MODBASE uses the Netscape plugin MODVIEW (V.A.Ilyin, U.Pieper, A.C.Stuart, M.A.Marti-Renom, L.McMahan and A.Sali, manuscript submitted for publication) to visualize and analyze the models of target sequences, template structures and their alignments. MODVIEW also contains a number of sequence and structure analysis tools. For example, it is possible to prepare multiple sequence alignments, multiple structure alignments, cluster protein sequences based on these alignments and study their variability. MODVIEW is currently available for the Linux operating system (http://guitar.rockefeller.edu/modview).

FUTURE DIRECTIONS

MODBASE will be updated continually to reflect the growth of the sequence and structure databases, as well as improvements in the methods and software used for calculating the models.

To facilitate the use of comparative protein structure models in classification of proteins and in annotation of their function, MODBASE will be integrated with additional resources in biology. In particular, we plan to link MODBASE to many major biological databases through the sequence retrieval system (SRS) (47). Similarly, we will integrate MODBASE into the distributed sequence annotation system (DAS) (http://stein.cshl.org/das/) (48).

MODBASE will also be expanded by adding additional sets of models, such as models for all single nucleotide polymorphisms and expressed sequence tags.

CITATION

Users of MODBASE are requested to cite this article in their publications.

ACKNOWLEDGEMENTS

We are especially grateful to Dr Roberto Sánchez for constructing the first version of MODBASE. We also thank Nebojsa Mirkovic, Bino John, William Lane, Maria Sammut and Edward Wittenstein for their contributions to MODBASE. This paper is based partly on publications by Baker and Sali (7) and Marti-Renom *et al.* (9). The project has been supported by NIH/GM R01 54762, NIH/GM P50 6M62529, Mathers

Foundation and Merck Genome Research Award. A.Stuart is an Alfred P. Sloan Postdoctoral Fellow. A.Sali is an Irma T. Hirschl Trust Career Scientist.

REFERENCES

1. Domingues,F.S., Koppensteiner,W.A. and Sippl,M.J. (2000) The role of protein structure in genomics. *FEBS Lett.*, **476**, 98–102.
2. Brenner,S.E. and Levitt,M. (2000) Expectations from structural genomics. *Protein Sci.*, **9**, 197–200.
3. Skolnick,J., Fetrow,J.S. and Kolinski,A. (2000) Structural genomics and its importance for gene function analysis. *Nat. Biotechnol.*, **18**, 283–287.
4. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated in this issue: *Nucleic Acids Res.* (2002), **30**, 245–248.
5. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
6. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18. Updated in this issue: *Nucleic Acids Res.* (2002), **30**, 17–20.
7. Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
8. Bonneau,R. and Baker,D. (2001) *Ab initio* protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 173–189.
9. 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.
10. Sanchez,R. and Sali,A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl Acad. Sci. USA*, **95**, 13597–13602.
11. Koehl,P. and Levitt,M. (1999) A brighter future for protein structure prediction. *Nature Struct. Biol.*, **6**, 108–111.
12. Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci.*, **10**, 352–361.
13. Eyrich,V., Marti-Renom,M.A., Przybylski,D., Fiser,A., Pazos,F., Valencia,A., Sali,A. and Rost,B. (2002) EVA: continuous automatic evaluation of structure prediction servers. *Bioinformatics*, in press.
14. Marti-Renom,M.A., Madhusudhan,M.S., Fiser,A., Rost,B. and Sali,A. (2002) Reliability of assessment of protein structure prediction methods. *Structure*, in press.
15. Johnson,M.S., Srinivasan,N., Sowdhamini,R. and Blundell,T.L. (1994) Knowledge-based protein modelling. *CRC Crit. Rev. Biochem. Mol. Biol.*, **29**, 1–68.
16. Xu,L.Z., Sanchez,R., Sali,A. and Heintz,N. (1996) Ligand specificity of brain lipid-binding protein. *J. Biol. Chem.*, **271**, 24711–24719.
17. Matsumoto,R., Sali,A., Ghildyal,N., Karplus,M. and Stevens,R.L. (1995) Packaging of proteases and proteoglycans in the granules of mast cells and other hematopoietic cells. A cluster of histidines on mouse mast cell protease 7 regulates its binding to heparin serglycin proteoglycans. *J. Biol. Chem.*, **270**, 19524–19531.
18. Irving,J.A., Whisstock,J.C. and Lesk,A.M. (2001) Protein structural alignments and functional genomics. *Proteins*, **42**, 378–382.
19. Luthy,R., Bowie,J.U. and Eisenberg,D. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.
20. Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.
21. Melon,F., Sánchez,R. and Sali,A. (2002) Statistical potentials for fold assessment. *Protein Sci.*, in press.
22. Fischer,D. and Eisenberg,D. (1997) Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl Acad. Sci. USA*, **94**, 11929–11934.
23. Guex,N., Diemand,A. and Peitsch,M.C. (1999) Protein modelling for all. *Trends Biochem. Sci.*, **24**, 364–367.
24. Rychlewski,L., Zhang,B. and Godzik,A. (1999) Functional insights from structural predictions: analysis of the *Escherichia coli* genome. *Protein Sci.*, **8**, 614–624.
25. Teichmann,S.A., Park,J. and Chothia,C. (1998) Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA*, **95**, 14658–14663.
26. Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.
27. Huynen,M., Doerks,T., Eisenhaber,F., Orengo,C., Sunyaev,S., Yuan,Y. and Bork,P. (1998) Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.*, **280**, 323–326.
28. 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.
29. Sali,A. (1998) 100,000 protein structures for the biologist. *Nature Struct. Biol.*, **5**, 1029–1032.
30. Burley,S.K., Almo,S.C., Bonanno,J.B., Capel,M., Chance,M.R., Gaasterland,T., Lin,D., Sali,A., Studier,F.W. and Swaminathan,S. (1999) Structural genomics: beyond the human genome project. *Nature Genet.*, **23**, 151–157.
31. Vitkup,D., Melamud,E., Moult,J. and Sander,C. (2001) Completeness in structural genomics. *Nature Struct. Biol.*, **8**, 559–566.
32. 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.
33. Sanchez,R. and Sali,A. (1999) ModBase: a database of comparative protein structure models. *Bioinformatics.*, **15**, 1060–1061.
34. 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.
35. 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.
36. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
37. Bonanno,J.B., Edo,C., Eswar,N., Pieper,U., Romanowski,M.J., Ilyin,V.A., Gerchman,S.E., Kycia,H., Studier,F.W., Sali,A. and Burley,S.K. (2001) Structural genomics of enzymes involved in sterol/isoprenoid biosynthesis. *Proc. Natl Acad. Sci. USA*, **98**, 12896–12901.
38. Gopal,S., Schroeder,M., Pieper,U., Sczyrba,A., Aytekin-Kurban,G., Bekiranov,S., Fajardo,J.E., Eswar,N., Sanchez,R., Sali,A. and Gaasterland,T. (2001) Homology-based annotation yields 1,042 new candidate genes in the *Drosophila melanogaster* genome. *Nature Genet.*, **27**, 337–340.
39. Burge,C.B. and Karlin,S. (1998) Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.*, **8**, 346–354.
40. Spahn,C.M.T., Beckmann,R., Eswar,N., Penczek,P., Sali,A., Blobel,G. and Frank,J. (2001) Structure of the 80S ribosome from *Saccharomyces cerevisiae*—tRNA–ribosome and subunit–subunit interactions. *Cell*, **107**, 373–386.
41. Jin,S., Martinek,S., Joo,W.S., Wortman,J.R., Mirkovic,N., Sali,A., Yandell,M.D., Pavletich,N.P., Young,M.W. and Levine,A.J. (2000) Identification and characterization of a p53 homologue in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA*, **97**, 7301–7306.
42. Bray,J.E., Todd,A.E., Pearl,F.M., Thornton,J.M. and Orengo,C.A. (2000) The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. *Protein Eng.*, **13**, 153–165.
43. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266. Updated in this issue: *Nucleic Acids Res.* (2002), **30**, 276–280.
44. Corpet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.
45. Stuart,A.C., Ilyin,V.A. and Sali,A. (2002) Ligbase: a database of families of aligned binding sites in known protein sequences and structures. *Bioinformatics*, in press.
46. Shindyalov,I.N. and Bourne,P.E. (2001) A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. *Nucleic Acids Res.*, **29**, 228–229.
47. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
48. Stein,L. (2001) Genome annotation: from sequence to biology. *Nature Rev. Genet.*, **2**, 493–503.