250–253 Nucleic Acids Research, 2000, Vol. 28, No. 1

MODBASE, a database of annotated comparative protein structure models

Roberto Sánchez, Ursula Pieper, Nebojša Mirkovic, Paul I. W. de Bakker, Edward Wittenstein and Andrej Šali*

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 1, 1999; Revised and Accepted October 11, 1999

ABSTRACT

MODBASE is a queryable database of annotated comparative protein structure models. The models are derived by MODPIPE, an automated modeling pipeline relying on the programs PSI-BLAST and MODELLER. The database currently contains 3D models for substantial portions of approximately 17 000 proteins from 10 complete genomes, including those of *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Escherichia coli*, as well as all the available sequences from *Arabidopsis thaliana and Homo sapiens*. The database also includes fold assignments and alignments on which the models were based. In addition, special care is taken to assess the quality of the models. ModBase is accessible through a web interface at http://guitar.rockefeller.edu/modbase/

INTRODUCTION

In a few years, the genome projects will provide us with the amino acid sequences of more than a million proteins—the catalysts, inhibitors, messengers, receptors, transporters and building blocks of the living organisms (1). The full potential of the genome projects will only be realized once we assign and understand the function of these proteins. While protein function is best determined experimentally, it can sometimes be predicted by matching the sequence of a protein with proteins of known function (2,3). Sequence-based predictions of function can be improved by considering three-dimensional (3D) structure of proteins (2,3). The 3D structure of a protein generally provides more information about its function than sequence alone because interactions of a protein with other molecules are determined by amino acid residues that are close in space even though they are frequently distant in sequence. In addition, because evolution tends to conserve function, which depends more directly on structure than on sequence, structure is more conserved in evolution than sequence. The net result is that patterns in space are frequently more recognizable than patterns in sequence.

Unfortunately, 3D structures have been determined for only a fraction of known protein sequences by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. While there are approximately 500 000 protein sequences in GenPept (4), there are only 10 000 experimentally determined protein structures in the Protein Data Bank (5; http://www.rcsb.org/pdb/). However, a useful 3D model can frequently be obtained by comparative or homology protein structure modeling, which constructs all-atom 3D models for those proteins that are related to at least one known protein structure (6,7).

The fraction of the known protein sequences that have at least one segment related to one or more known structures varies with a genome, currently ranging from 20 to 50% (8–15). 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. Furthermore, the fraction of protein sequences that can be modeled reliably by comparative modeling is increasing rapidly. It has been estimated that globular protein domains cluster in only a few thousand fold families. Approximately 900 of these folds have already been structurally defined (16–18). Assuming the current growth rate in the number of known protein structures, the structure of at least one member of most globular folds will be determined in less than 10 years (19). Structural genomics may in fact accelerate this goal (20–25). As a result, comparative modeling will become applicable to most of the globular protein domains soon after the completion of the human genome project.

Two examples of comparative modeling for complete genomes have already been described (10,26), demonstrating that it is possible to automate comparative modeling for large-scale applications. Despite the usefulness of comparative modeling, it is still not a common sequence analysis tool for the biologist, partly due to the lack of easy access to reliable and evaluated models. The database described in this paper attempts to resolve this problem.

CONTENTS

The database currently contains models for segments of approximately 17 000 proteins from the completely sequenced genomes of *Saccharomyces cerevisiae*, *Mycoplasma genitalium*, *Caenorhabditis elegans*, *Escherichia coli*, *Methanobacterium thermoautotrophicum*, *Synechocystis* sp., *Pyrococcus horikoshii*, *Methanococcus jannaschii*, *Haemophilus influenzae* and *Mycoplasma pneumoniae*, as well as all *Arabidopsis thaliana*

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

and *Homo sapiens* proteins in the SWISS-PROT database (27). The sources of the genomes are listed at http://guitar.rockefeller.edu/modbase/sources.html . Each model has its non-hydrogen atom coordinates stored in a flat file in PDB format. The format of the files allows for inclusion of information about the modeling process (A.Adzhubei, N.Guex and M.Peitsch, unpublished). The database also contains all fold assignments, alignments and model evaluations.

Models are generated with an entirely automated four-step procedure implemented in the MODPIPE pipeline software (10,28): (i) fold assignment, (ii) sequence–structure alignment, (iii) model building, and (iv) model evaluation. The procedure can be applied independently and in parallel on a cluster of workstations to thousands of protein sequences, including complete genomes and large protein sequence databases. For fold assignment, each sequence from a genome is compared with a non-redundant set of proteins of known 3D structure using PSI-BLAST (29). Next, for each target protein sequence, a multiple global alignment with the matching structures is constructed by the ALIGN2D command in the program MODELLER (30). This alignment tends to be more accurate than the PSI-BLAST alignment because (i) it includes all the sequences and structures that are sufficiently similar to the target sequence, (ii) it uses a structure-dependent gap penalty function to position gaps in a structurally reasonable environment, and (iii) it matches complete structural domains as obtained from the known template structures (R.Sánchez, F.Melo, N.Mirkovic and A.Šali, in preparation). In the third step, the sequence–structure alignment is used to build a 3D model for the matched parts of the target protein sequence by the program MODELLER. Finally, the model is evaluated as discussed next.

Model evaluation is essential for assessing the value of 3D protein models in any protein structure prediction (7,31,32). It is especially important for MODPIPE because a relatively permissive cutoff is used to select known protein structures for model building in the first fold assignment step. This permissivness reduces the number of missed hits, but it also increases the number of false fold assignments and alignment mistakes. The fold assignment errors begin to appear when relatively dissimilar template–target sequences are matched (i.e., <30% sequence identity). In addition, even if the fold is assigned correctly, errors in the alignment may still result in a bad model. The alignment errors can be significant when the sequence identity drops below 35%. A reliable model is obtained only if both the correct fold assignment and an approximately correct alignment are made.

The overall accuracy of a model is measured by an overlap between the model and the actual structure. The overlap is defined as the fraction of residues whose Cα atoms are within 3.5 Å of each other in the globally superposed pair of structures. Models that overlap with the correct structures in >30% of their residues are defined here as 'good' models. Such models are likely to have a correct fold, which is frequently sufficient for coarse prediction of protein function (33). A method for calculating the probability of whether a given model is good, pG, was developed (10) and is used to evaluate all the models in MODBASE. If a given model has pG > 0.5, it is called a 'reliable' model. The method depends on a statistical scoring function (32) and was calibrated using 3993 and 6270 good and bad models for 1085 proteins of known structure (10). An assessment of the method by the jack-knife procedure indicated that for models longer than 100 residues the classification results in <5% of false positives and <8% of false negatives.

Combined 3D modeling and model evaluation is the best way of either confirming or rejecting a match between remotely related sequence and structure (10,34). This is important because most of the related protein pairs share <30% sequence identity (10). As a result MODBASE includes reliable models based on templates that are not detectable as significant matches by PSI-BLAST alone.

ACCESS AND INTERFACE

MODBASE has a web interface at http://guitar.rockefeller.edu/modbase/ . Models for yeast proteins are also accessible through links from the Sacch3D (35) database at http://genome-www.stanford.edu/Sacch3D . The database is searchable by SWISS-PROT/TrEMBL and GenPept accession numbers, as well as by ORF names, keywords, model reliability, model size, target–template sequence identity and alignment significance (Fig. 1a). It is also possible to perform sequence similarity searches against the model sequences using BLAST (29). Searching results in a table of models satisfying all search criteria (Fig. 1b). The table lists the modeled regions, the templates used to construct the models, target–template similarities and model reliabilities. For each model, it also includes links to a more detailed description of the model, to a summary of all models for a given protein, and to the PDB for a detailed description of the template structure used in modeling. If the modeled sequence is present in SWISS-PROT/TrEMBL, its description is displayed together with a link to the database. The model description page contains a graphical representation of the target–template alignment (Fig. 1c). In addition, it is linked to the model coordinates in the PDB format, to the target–template alignment used to derive the model (Fig. 1c), and to a display of the model by the 3D visualization program Rasmol (36) (Fig. 1d). The model description page also contains links to the MODBASE entries related to the target sequence and to the CATH domains (17) contained in the model. Finally, statistical data, such as distributions of several model properties in MODBASE can also be displayed.

USING COMPARATIVE MODELS

It is frequently possible to extract more information from a comparative model than from the modeled sequence alone, or even from its alignment to a related protein structure (7,28). For example, the preferred ligand of brain lipid binding protein could be predicted correctly from the volume and shape of the ligand binding cleft in its comparative model (37). Another example is provided by mouse mast cell proteases, some of which have a conserved surface region of positively charged residues that binds proteoglycans (38). This region is not easily recognizable in the sequence or its alignment to a known structure because the constituting residues occur at variable and sequentially non-local positions in sequence that form a binding site only when the protease is fully folded.

In general, comparative modeling has been applied successfully to many biological problems (6,7). It can be helpful in proposing and testing hypotheses in molecular biology, such as hypotheses about ligand binding sites (37,38), substrate specificity

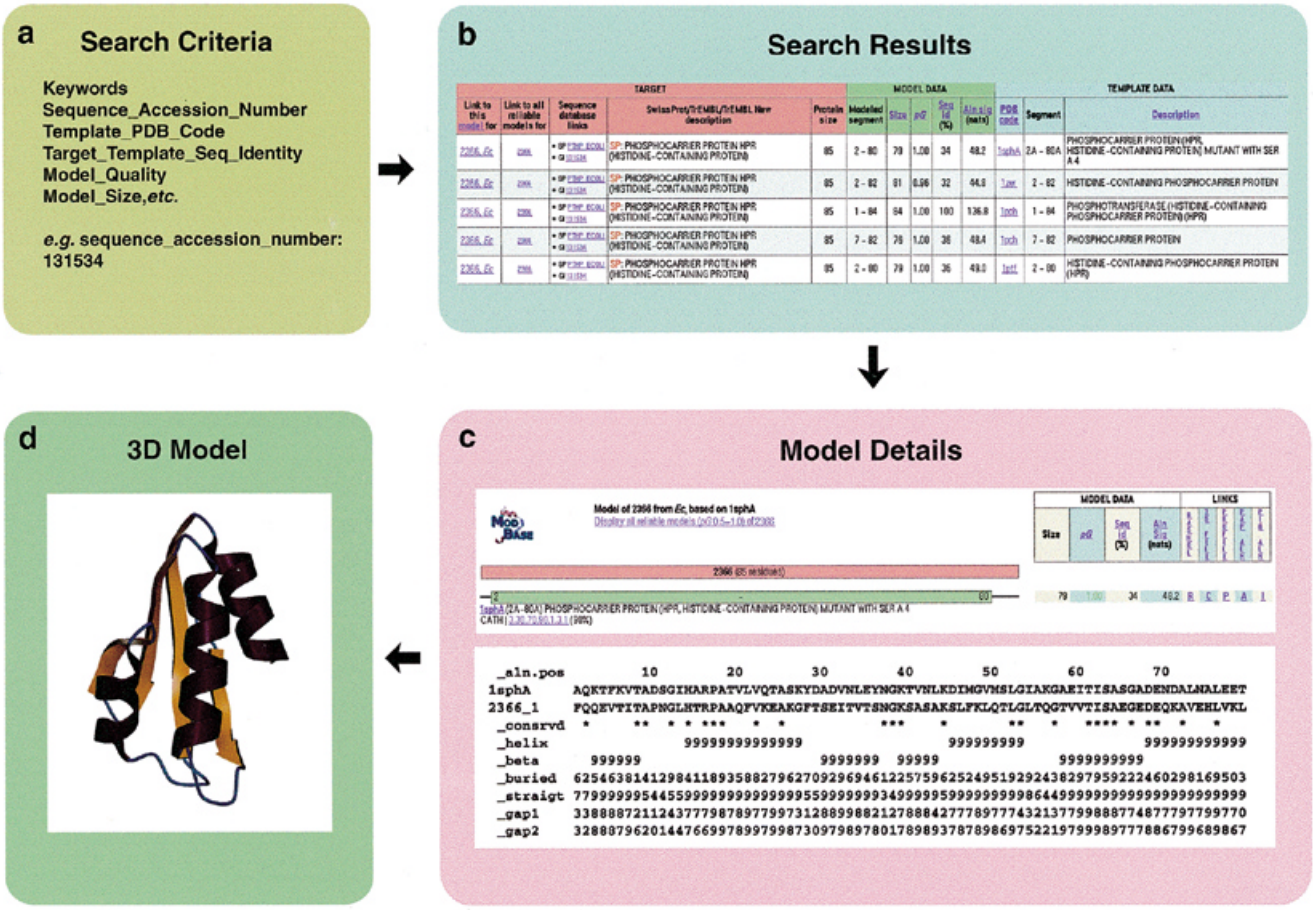

Figure 1. The contents of MODBASE. See text for details.

(39), drug design (40) and protein–protein interactions (41). It can also provide starting models in X-ray crystallography (42) and NMR spectroscopy (43). Another use of 3D models is that some binding and active sites, which cannot possibly be found by searching for local sequence patterns, frequently should be detectable by searching for small 3D motifs that are known to bind or act on specific ligands (44–46). Finally, comparative models in combination with model evaluation can also be used to confirm or reject remote sequence–structure relationships, complementing the existing sequence matching and threading methods for fold assignment (10,34).

FUTURE DIRECTIONS

The fraction of protein sequences that can be modeled with useful accuracy by comparative modeling is increasing rapidly. The main reasons for this improvement are the increased numbers of known folds and the structures per fold family (19), as well as improvements in the fold assignment and comparative modeling techniques (47,48). For example, potential enhancements of coverage and model quality in MODBASE include the use of more sophisticated fold assignment and alignment methods, such as threading sequences through structures (49) and relying on many homologous sequences at the same time to construct Hidden Markov Models (50).

In the future, MODBASE will grow to reflect (i) the growth of the sequence databases, (ii) the growth of the database of known protein structures, (iii) and improvements in the software for calculating the models. It is expected that the SWISS-PROT+TrEMBL protein sequence databases and various EST databases will be processed soon.

CITATION

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

ACKNOWLEDGEMENTS

We are grateful to Dr Steve A. Chervitz for making links from Sacch3D to MODBASE, and to Dr Ashley Stuart for comments on the manuscript. R.S. is a Howard Hughes Medical Institute predoctoral fellow. A.Š. is a Sinsheimer Scholar and an Alfred P. Sloan Research Fellow. The project has also been aided by grants from NIH (GM 54762) and NSF (BIR-9601845).

REFERENCES

1. Collins,F.S., Patrinos,A., Jordan,E., Chakravarti,A., Gesteland,R. and Walters,L. (1998) *Science*, **282**, 682–689.
2. Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) *J. Mol. Biol.* **283**, 707–725.
3. Koonin,E.V., Tatusov,R.L. and Galperin,M.Y. (1998) *Curr. Opin. Struct. Biol*., **3**, 355–363.
4. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F.F., Rapp,B.A. and Wheeler,D.L. (1999) *Nucleic Acids Res*. **27**, 12–17. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 15–18.
5. Abola,E.E., Bernstein,F.C., Bryant,S.H., Koetzle,T. and Weng,J. (1987) In Allen,F.H., Bergerhoff,G. and Sievers,R. (eds), *Crystallographic Databases Information, Content, Software systems, Scientific applications*. Data Commission of the International Union of Crystallography Bonn/Cambridge/Chester, pp. 107–132.
6. Johnson,M.S., Srinivasan,N., Sowdhamini,R. and Blundell,T.L. (1994) *CRC Crit. Rev. Biochem. Mol. Biol*., **29**, 1–68.
7. Martí-Renom,M.A., Stuart,A., Fiser,A., Sánchez,R., Melo,F. and Šali,A. (2000) *Annu. Rev. Biophys. Biomol. Struct.*, in press.
8. Guex,N., Diemand,A. and Peitsch,M.C. (1999) *Trends Biochem. Sci*., **24**, 364–367.
9. Fischer,D. and Eisenberg,D. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 11929–11934.
10. Sanchez,R. and Sali,A. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 13597–13602.
11. Rychlewski,L., Zhang,B. and Godzik,A. (1998) *Fold. Des.*, **3**, 229–238.
12. Huynen,M., Doerks,T., Eisenhaber,F., Orengo,C., Sunyaev,S., Yuan,Y. and Bork,P. (1998) *J. Mol. Biol.* **280**, 323–326.
13. Grandori,R. (1998) *Protein Eng*., **11**, 1129–1135.
14. Teichmann,S.A., Park,J. and Chothia,C. (1998) *Proc. Natl Acad. Sci. USA*, **22**, 14658–14663.
15. Jones,D.T. (1999) *J. Mol. Biol.*, **287**, 797–815.
16. Hubbard,T.J.P., Ailey,B., Brenner,S.E., Murzin,A.G. and Chothia,C. (1999) *Nucleic Acids Res*., **27**, 254–256. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 257–259.
17. Orengo,C.A., Pearl,F.M.G., Bray,J.E., Todd,A.E., Martin,A.C., Conte,L.L. and Thornton,J.M. (1999) *Nucleic Acids Res*., **27**, 275–279. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 277–282.
18. Holm,L. and Sander,C. (1999) *Nucleic Acids Res*., **27**, 244–247.
19. Holm,L. and Sander,C. (1996) *Science*, **273**, 595–602.
20. Terwilliger,T.C., Waldo,G., Peat,T.S., Newman,J.M., Chu,K. and Berendzen,J. (1998) *Protein Sci.*, **7**, 1851–1856.
21. Šali,A. (1998) *Nature Struct. Biol.*, **5**, 1029–1032.
22. Zarembinski,T.I., Hung,L.W., Mueller-Dieckmann,H.J., Kim,K.K., Yokota,H., Kim,R. and Kim,S.H. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 15189–15193.
23. Burley,S.K., Almo,S.C., Bonanno,J.B., Capel,M., Chance,M.R., Gaasterland,T., Lin,D., Šali,A., Studier,F.W. and Swaminathan,S. (1999) *Nature Genet.*, **23**, 151–157.
24. Montelione,G.T. and Anderson,S. (1999) *Nature Struct. Biol.*, **6**, 11–12.
25. Cort,J.R., Koonin,E.V., Bash,P.A. and Kennedy,M.A. (1999) *Nucleic Acids Res*., **27**, 4018–4027.
26. Peitsch,M.C., Wilkins,M.R., Tonella,L., Sanchez,J.C., Appel,R.D. and Hochstrasser,D.F. (1997) *Electrophoresis*, **18**, 498–501.
27. Bairoch,A. and Apweiler,R. (1999) *Nucleic Acids Res*., **27**, 49–54. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 45–48.
28. Sánchez,R. and Sali,A. (1999) *J. Comp. Phys.*, **151**, 388–401.
29. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res*., **25**, 3389–3402.
30. Šali,A. and Blundell,T.L. (1993) *J. Mol. Biol.*, **234**, 779–815.
31. Lüthy,R., Bowie,J.U. and Eisenberg,D. (1992) *Nature*, **356**, 83–85.
32. Sippl,M.J. (1993) *Proteins*, **17**, 355–362.
33. Orengo,C.A., Jones,D.T. and Thornton,J.M. (1994) *Nature*, **372**, 631–634.
34. Guenther,B., Onrust,R., Sali,A., O'Donnell,M. and Kuriyan,J. (1997) *Cell*, **91**, 335–345.
35. Chervitz,S.A., Hester,E.T., Ball,C.A., Dolinski,K., Dwight,S.S., Harris,M.A., Juvik,G., Malekian,A., Roberts,S., Roe,T., Scafe,C., Schroeder,M., Sherlock,G., Weng,S., Zhu,Y., Cherry,J.M. and Botstein,D. (1999) *Nucleic Acids Res.*, **27**, 74–78. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 77–80.
36. Sayle,R. and Milner-White,E.J. (1995) *Trends Biochem. Sci.*, **20**, 374.
37. Xu,L.Z., Sánchez,R., Šali,A. and Heintz,N. (1996) *J. Biol. Chem.*, **271**, 24711–24719.
38. Matsumoto,R., Šali,A., Ghildyal,N., Karplus,M. and Stevens,R.L. (1995) *J. Biol. Chem.*, **270**, 19524–19531.
39. Caputo,A., James,M.N.G., Powers,J.C., Hudig,D. and Bleackley,R.C. (1994) *Nature Struct. Biol.*, **1**, 364–367.
40. Ring,C.S., Sun,E., McKerrow,J.H., Lee,G.K., Rosenthal,P.J., Kuntz,I.D. and Cohen,F.E. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 3583–3587.
41. Vakser,I.A. (1997) *Proteins Suppl.*, **1**, 226–230.
42. Carson,M., Bugg,C.E., Delucas,L. and Narayana,S. (1994) *Acta Crystallogr*., D**50**, 889–899.
43. Nagata,T., Gupta,V., Kim,W.-Y., Šali,A., Chait,B.T., Shigesada,K., Ito,Y. and Werner,M.H. (1999) *Nature Struct. Biol.*, **6**, 615–619.
44. Wallace,A., Borkakoti,N. and Thornton,J.M. (1997) *Protein Sci.*, **6**, 2308–2323.
45. Fetrow,J.S. and Skolnick,J. (1998) *J. Mol. Biol.*, **281**, 949–968.
46. Kleywegt,G.J. (1999) *J. Mol. Biol.*, **285**, 1887–1897.
47. Dunbrack,R.L.,Jr, Gerloff,D.L., Bower,M., Chen,X., Lichtarge,O. and Cohen,F.E. 1997) *Folding Des.*, **2**, R27–R42.
48. Koehl,P. and Levitt,M. (1999) *Nature Struct. Biol.*, **6**, 108–111.
49. Jones,D. (1997) *Curr. Opin. Struct. Biol.*, **7**, 377–387.
50. Eddy,S.R. (1996) *Curr. Opin. Struct. Biol.*, **6**, 361–365.