BIOINFORMATICS APPLICATIONS NOTE Vol. 15 no. 12 1999
Pages 1060–1061

# MODBASE: *A database of comparative protein structure models*

*Roberto Sánchez and Andrej Šali*

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

Received on March 16, 1999; revised on June 9, 1999; accepted on June 24, 1999

**Abstract**
***Summary:*** MODBASE *is a database of evaluated and annotated comparative protein structure models. The database also includes fold assignments and alignments on which the models were based.*
***Availability:*** MODBASE *is accessible on the Web at http://guitar.rockefeller.edu/modbase. Models for yeast proteins are also accessible through links from the* SACCH3D *database at http://genome-www.stanford.edu/Sacch3D.*
***Contact:*** *sali@rockefeller.edu; http//guitar.rockefeller.edu/*

Native three-dimensional structure (3D) of a protein is valuable in testing, understanding, and modifying protein function. While 3D structures of only a tiny fraction of known protein sequences (Benson *et al.*, 1999) have been defined experimentally (Abola *et al.*, 1987), comparative modeling can frequently provide a useful 3D model of a protein (Johnson *et al.*, 1994; Sánchez and Šali, 1997b). 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 SWISS-MODEL (Guex *et al.*, 1999) database of comparative models attempts to resolve this problem, as does the MODBASE database described in this paper.

MODBASE is a database of annotated comparative protein structure models. The models consist of coordinates for all non-hydrogen atoms in the modeled part of a protein. Models are generated entirely automatically in a four step procedure (Sánchez and Šali, 1998, 1999): (i) fold assignment, (ii) sequence–structure alignment, (iii) model building, and (iv) model evaluation. This procedure can be applied to thousands of protein sequences, including complete genomes and large protein sequence databases. In the fold assignment step, each sequence from a genome is compared with a non-redundant set of proteins of known 3D structure (Abola *et al.*, 1987). This is achieved by an iterative sequence similarity search by program PSI-BLAST (Altschul *et al.*, 1997). In the second step, the matching parts of a given protein sequence and a related known protein structure are aligned by the ALIGN2D command of MODELLER (Sánchez and Šali, in preparation). This procedure places gaps in the structurally reasonable context. In the third step, all the pairwise sequence–structure alignments are used individually to build 3D models for the matched parts of the protein sequences by the program MODELLER (Šali and Blundell, 1993; Sánchez and Šali, 1997a). The fourth step, evaluation of models, is discussed in the following section.

It is essential for assessing the value of 3D protein models to estimate their overall accuracy (Lüthy *et al.*, 1992; Sippl, 1993; Sánchez and Šali, 1997b). In the fold assignment step of the pipeline, a relatively permissive cutoff is used for selecting known protein structures for model building. This results in a smaller number of missed hits, but it also increases the number of false fold assignments and the number of mistakes in alignments. 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 $C_\alpha$ atoms are within 3.5 Å of each other in the globally superposed pair of structures. Models that overlap with the correct structures in more than 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 (Orengo *et al.*, 1994). A method for calculating the probability of whether a given model is good, $pG$, was developed (Sánchez and Šali, 1998) and is used to evaluate all the models in MODBASE.

The database currently contains models for segments of more than 17,000 proteins in *Saccharomyces cerevisiae*, *Mycoplasma genitalium*, *Caenorhabditis elegans*, *Es-*

**Table 1.** Contents of ModBase

| Organism | Proteins with models[a] | Models[b] | % of organism proteins with models | % of organism residues modeled |
|---|---|---|---|---|
| *Saccharomyces cerevisiae* | 2587 | 4484 | 42 | 20 |
| *Mycoplasma genitalium* | 216 | 280 | 45 | 29 |
| *Caenorhabditis elegans* | 7900 | 13523 | 39 | 22 |
| *Escherichia coli* | 1625 | 2560 | 38 | 27 |
| *Methanobacterium thermo.* | 663 | 1125 | 21 | 19 |
| *Synechocystis* sp. | 1000 | 1670 | 38 | 25 |
| *Pyrococcus horikoshii* | 611 | 946 | 30 | 24 |
| *Methanococcus jannaschii* | 630 | 987 | 36 | 28 |
| *Haemophilus influenzae* | 670 | 1217 | 40 | 30 |
| *Aquifex acolicus* | 665 | 1063 | 44 | 31 |
| *Mycoplasma pneumoniae* | 244 | 297 | 18 | 16 |
| *Sulfolobus solfataricus* | 301 | 579 | 30 | 26 |

[a]The number of proteins that have at least one segment modeled reliably. Whether or not a model is reliable is predicted as described briefly in the text, and in more detail in Sánchez and Šali (1998).
[b]The number of models calculated for the genome. This number is larger than the number of proteins modeled because many proteins have independently calculated models for the same domain in the protein, as well as independently calculated models for different domains in the same protein.

*cherichia coli*, *Methanobacterium thermoautotrophicum*, *Synechocystis* sp., *Pyrococcus horikoshii*, *Methanococcus jannaschii*, *Haemophilus influenzae*, *Aquifex aeolicus*, *Mycoplasma pneumoniae* and *Sulfolobus solfataricus* (Table 1).

The database is searchable by protein names, keywords, template structure, organism, model reliability, model size, target–template sequence identity, and alignment significance. It is also possible to search for sequence similarities to the model sequences using BLAST (Altschul *et al.*, 1997). Searching produces a table of models satisfying all search criteria. The table lists the modeled regions of the target proteins, 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, a summary of all models for a given protein, and the PDB database (Abola *et al.*, 1987) for a detailed description of the template structure used in modeling. The model description page contains a schematic representation of the target-template alignment and links to the template fold entries in the CATH database (Orengo *et al.*, 1999). In addition, it links to the model coordinates in the PDB format, the target-template alignment used to derive the model, and display of the model by the 3D visualization program RASMOL (Sayle and Milner-White, 1995).

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 (Bairoch and Apweiler, 1999) and various EST databases will be processed by the end of 1999.

## Acknowledgments

We are grateful to Dr. Steve A. Chervitz for making links from SGD to ModBase and Paul de Bakker for help in implementing the WWW interface. RS is a Howard Hughes Medical Institute predoctoral fellow. AŠ 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

Abola,B.B., Bernstein,F.C., Bryant,S.H., Koetzle,T. and Weng,J. (1987) Protein data bank. 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.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.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.

Bairoch,A. and Apweiler,R. (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.*, **27**, 49–54.

Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Guellette,B.F. F., Rapp,B.A. and Wheeler,D.L. (1999) Genbank. *Nucleic Acids Res.*, **27**, 12–17.

Guex,N., Diemand,A. and Peitsch,M.C. (1999) Protein modelling for all. *Trends Biochem. Sci.*, **24**, 364–367.

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.

Lüthy,R., Bowie,J.U. and Eisenberg,D. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.

Orengo,C.A., Jones,D.T. and Thornton,J.M. (1994) Protein superfamilies and domain super-folds. *Nature*, **372**, 631–634.

Orengo,C.A., Pearl,F.M. G., Bray,J.B., Todd,A.B., Martin,A.C., Conte,L.L. and Thornton,J.M. (1999) The CATH database provides insights into protein structure/function relationship. *Nucleic Acids Res.*, **27**, 275–279.

Šali,A. and Blundell,T.L. (1993) Comparative protein modelling bysatisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Sánchez,R. and Šali,A. (1997a) Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins*, **Suppl. 1**, 50–58.

Sánchez,R. and Šali,A. (1997b) Advances in comparative protein-structure modeling. *Curr. Opin. Struct. Biol.*, **7**, 206–214.

Sánchez,R. and Šali,A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl Acad. Sci. USA*, **95**, 13597–13602.

Sánchez,R. and Šali,A. (1999) Comparative protein structure modeling in genomics. *J. Comp. Phys.*, **151**, 388–401.

Sayle,R. and Milner-White,B.J. (1995) RasMol: Biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.

Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.