Nucleic Acids Research, 2006, Vol. 34, Database issue ***D315–D318***
doi:10.1093/nar/gkj056

The SWISS-MODEL Repository: new features and functionalities

Jürgen Kopp and Torsten Schwede*

Biozentrum der Universität Basel and Swiss Institute of Bioinformatics, Basel, Switzerland

Received September 15, 2005; Accepted October 4, 2005

ABSTRACT

The SWISS-MODEL Repository is a database of annotated 3D protein structure models generated by the SWISS-MODEL homology-modelling pipeline. As of September 2005, the repository contained 675 000 models for 604 000 different protein sequences of the UniProt database. Regular updates ensure that the content of the repository reflects the current state of sequence and structure databases, integrating new or modified target sequences, and making use of new template structures. Each Repository entry consists of one or more 3D models accompanied by detailed information about the target protein and the model building process: functional annotation, a detailed template selection log, target-template alignment, summary of the model building and model quality assessment. The SWISS-MODEL Repository is freely accessible at http://swissmodel.expasy.org/repository/.

INTRODUCTION

The rational design of many types of biological experiments is greatly facilitated by insights into the molecular mechanisms of biological macromolecules gained from their 3D structures. Proteins as the 'working molecules' of living cells are in the focus of several systematic structural genomics efforts (1). Although tremendous progress has been made in experimental structure determination by X-ray crystallography and NMR in recent years, the number of structurally characterized proteins is still small compared with the number of known protein sequences: while the UniProt protein knowledge base (2) reported more than 2 million sequence entries in September 2005, at the same time only ~35 000 database entries for experimentally determined protein structures were deposited in the Protein Data Bank (PDB) (3).

A wide spectrum of computational methods to predict the 3D structures of proteins has been developed to overcome this limitation, ranging from *ab initio* methods aiming to predict protein structures without prior knowledge, to comparative modelling approaches making use of information from experimentally determined protein structures. As the 3D structures of proteins are better conserved than their sequences (4), the experimentally known structure of one protein can be used as template to generate a model for other proteins of the same family sharing sufficient sequence homology (target). The accuracy and reliability of different structure prediction approaches have been assessed in several rounds of CASP (Critical Assessment of Techniques for Protein Structure Prediction) experiments (5). Comparative (or homology) modelling has been shown to be the most accurate method (6), able to reliably build 3D protein models with sufficient accuracy for applications such as structure-based drug design, or rationalizing the effects of sequence variations (7–9).

All homology modelling approaches consist of the following four steps: identification of suitable template structure(s), alignment of template structure(s) and target sequence, building of one or more 3D model(s), and model quality evaluation. These steps can be repeated iteratively until a satisfying modelling result is achieved and the best model is selected (7,10).

The number of structurally uncharacterized protein sequences is huge and growing constantly, and at the same time new template structures are being determined by individual research groups or structural genomics efforts (1). Fully automated, stable and reliable modelling pipelines (11–13) had to be created to handle this steady flow of new data. SWISS-MODEL has been established as an expert system for automated homology modelling, integrating the necessary databases, tools and computing resources in an easy to use web-based modelling workbench (11,14). The accuracy and reliability of the SWISS-MODEL server are continuously monitored by the EVA-CM evaluation project (15).

Interactive modelling servers require significant time to complete a certain modelling request, whereas model databases provide instant and queryable access to pre-computed models. Databases of models generated by automated procedures on large scale, such as SWISS-MODEL Repository (16) or ModBase (17), give easy access to regularly updated homology models and thereby help to enrich other database projects

*To whom corresponding should be addressed. Tel: +41 61 267 15 81; Fax: +41 61 267 15 84; Email: Torsten.Schwede@unibas.ch

with structural information. Structural genomics projects and model databases complement one another in the coverage of the protein structure space (18). In this paper we describe new features and improved functionalities of the SWISS-MODEL Repository.

SWISS-MODEL REPOSITORY

The SWISS-MODEL Repository is a database of annotated 3D protein models created by the fully automated SWISS-MODEL server pipeline (14). The model database can be accessed via a web interface, and the models are freely available at http://swissmodel.expasy.org/repository/. For a given UniProt entry, a model navigator indicates those regions in the protein, where 3D models are available and allows for fast selection of the individual models. The navigator has been extended to incorporate biological annotations via InterPro classification (19), and, in combination with the Target Sequence Info view, provides the user with detailed information about the target sequence. The model section (Figure 1a) presents a summary for each individual model, including information about the template structure, and the target-template alignment used for model building. Detailed modelling and template selection logs contain essential information about the modelling process, e.g. handling of insertions and deletions. As quality evaluation is indispensable for a predictive method like homology modelling, graphical representations of model assessment by Anolea mean force potentials (20) and Gromos force field energy (21) are given for each model (Figure 1d).

Database content

The number of models in the repository has more than doubled over the last two years. As of September 2005, the SWISS-MODEL Repository contained 675 000 protein models for 604 000 distinct UniProt (Release 5.8) entries. For 75 150 (39%) SwissProt entries (Release 47.8) and 528 932 (27%) TrEMBL (Release 30.8) entries a significant part of the sequence could be modelled. This corresponds to a per-residue coverage of 28% for SwissProt and 17% for TrEMBL. Model sizes range from 44 residues for the shortest model to 1512 residues for the longest model (ferredoxin-dependent glutamate synthase from *Anabena* sp.), with an average model size of 200 amino acids. Of all models 25% correspond to bacterial proteins, 1% to archaea, 28% to viruses and 46% to eukaryotic proteins, including 4.3% human sequences. The aim of the SWISS-MODEL Repository is to provide accurate models suitable for the design of biological experiments. Therefore, care was taken to include only models with a high level of confidence, i.e. reliable target-template alignments with a sequence identity higher than 40%, and acceptable evaluation results by Anolea (20) and Gromos (21).

Cross-references with other databases

The SWISS-MODEL Repository is cross-referenced with UniProt and InterPro. A subset of high quality models of the repository is complementing the structural information available in these two databases. Each entry in the repository contains a sequence-based CRC64 checksum, which uniquely identifies its corresponding amino acid sequence and guarantees data consistency when exchanging records between different databases. Moreover, this mechanism can also ensure that hyperlinks between websites using UniProt accession codes are referencing identical protein sequences. Therefore hyperlinks to the SWISS-MODEL Repository consist of the URL, a UniProt accession code, and optionally the start residue of a model, and the CRC64 checksum of the amino acid sequence.

Visualization of functional annotation

The interpretation of biological annotations of a protein sequence in the context of its 3D structure is of great help for a detailed understanding of its molecular function. Each target sequence in the repository has a short protein description with links to relevant sequence based resources. A graphical representation of the InterPro functional and domain level annotations of the target sequence indicates regions of functional and structural importance in the target sequence (Figure 1b). Their location relative to the available models can be visualized directly from the website using the interactive Java based AstexViewer applet (22) enabling the user to analyse biologically significant substructures. Alternatively, model coordinates can be downloaded for visualization (Figure 1c), and further analysis using specialized software, such as DeepView (11), Dino (http://www.dino3d.org) or RasMol (23).

Incremental update procedure

Computing an entire release of the repository takes several weeks on our in-house Linux cluster and is done when major changes have been implemented, e.g. algorithmic improvements in the pipeline. Fast integration of new sequence data and new experimentally determined protein structures released by the PDB (3) is a big benefit for the user, but poses a challenge to the update capability of the model database. We therefore established an incremental update procedure which takes into account new template structures, new protein sequence entries and modified sequences in UniProt. New or modified protein sequences are modelled and incorporated into the repository. When new experimentally determined structures become available, models are generated for those proteins, which could not be modelled previously, and existing models are updated for those proteins, for which better templates are identified among the newly released structures. Current data status of the SWISS-MODEL Repository is shown on the entry page.

Template selection

The automated modelling pipeline performs a sequence based search of the SWISS-MODEL template library to select the best template to model a specific region of the target sequence, favouring template structures with high resolution (14). However, the template selection log lists all template structures identified during the template search with links to the SWISS-MODEL template library. For each putative template, information about experimental details, likely quaternary structure (24), bound ligands, and links to external structure databases, such as SCOP (25), CATH (26) and MSD (27), are provided accompanied by a small ribbon representation. This enables the user to select alternative templates, e.g. templates

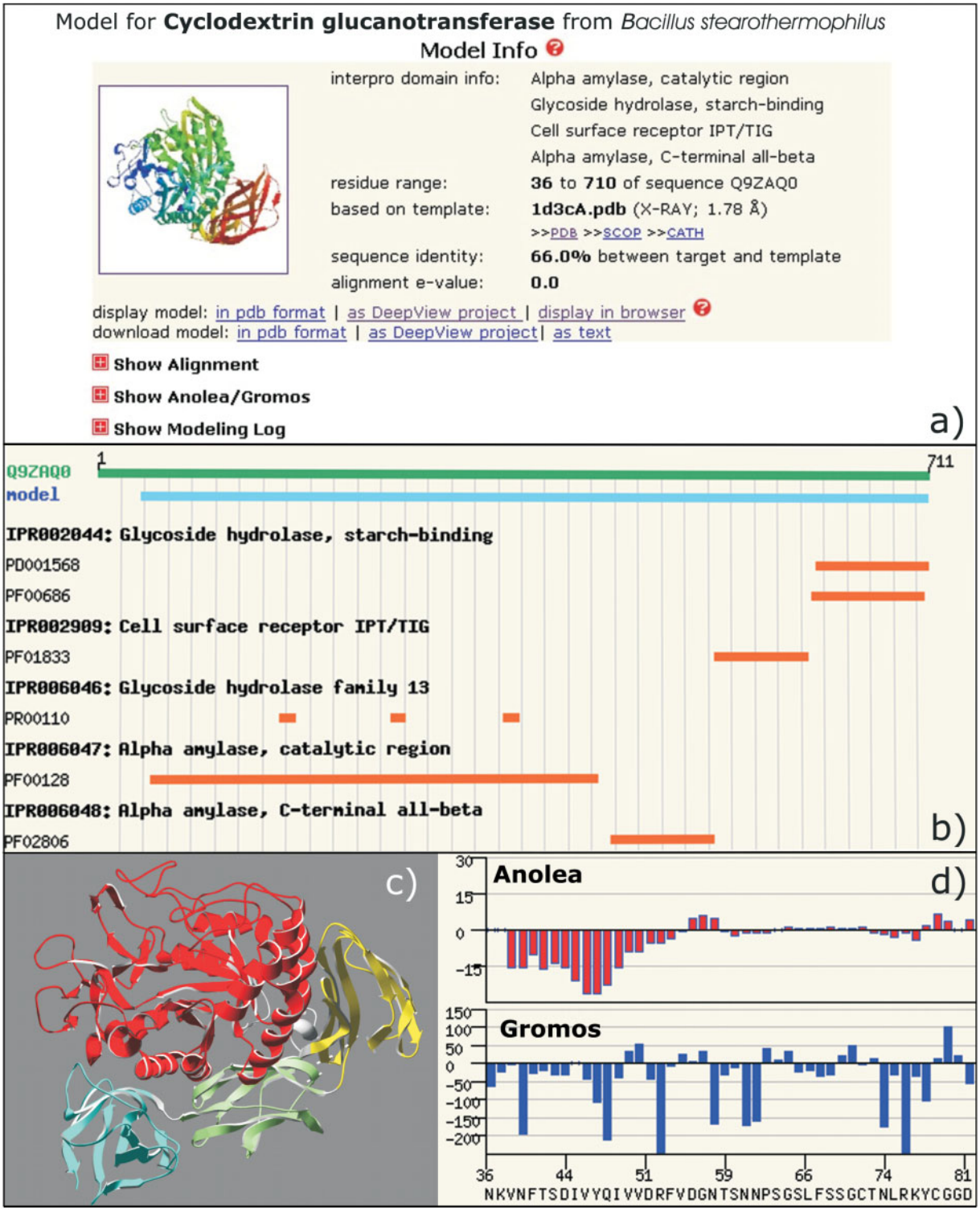

Figure 1. SWISS-MODEL Repository entry for Cyclodextrin glucanotransferase from *Bacillus stearothermophilus* (UniProt: Q9ZAQ0). (**a**) The 'Model Info' section provides a description of the template structure, the target-template alignment and a log of the modelling process. (**b**) Annotation of functional sites and individual domains of the target protein using InterPro assignments. (**c**) Structure of the model as ribbon representation using DeepView/PovRay coloured according to InterPro assignments. (**d**) Model quality evaluation are provided as Anolea and Gromos plots.

with or without ligand, open or closed conformations, multimeric structures, which might be more appropriate for the specific modelling task, and to continue the modelling project using the interactive SWISS-MODEL server architecture.

ACKNOWLEDGEMENTS

We are deeply indebted to Manuel C. Peitsch (Novartis AG, Basel, Switzerland) and to Nicolas Guex (GSK, Raleigh, NC)

for their pioneering work on large-scale protein structure modelling. We are grateful to Nicola Mulder, Rolf Apweiler (EBI Hinxton, UK), Isabelle Phan and Amos Bairoch (SIB Geneva, Switzerland) for the fruitful collaboration. We thank Lorenza Bordoli (EMBnet and Biozentrum and SIB Basel) for very encouraging discussions and committed user support. Aspects of the integration of SWISS-MODEL Repository and InterPro have been funded by the European Union under grant QLRI-CT-2001-00015 under the RTD program 'Quality of Life and Management of Living Resources'. We would like to acknowledge the financial support by the Swiss National Science Foundation (SNF). Funding to pay the Open Access publication charges for this article was provided by the Swiss Institute of Bioinformatics.

Conflict of interest statement. None declared.

REFERENCES

1. Todd,A.E., Marsden,R.L., Thornton,J.M. and Orengo,C.A. (2005) Progress of structural genomics initiatives: an analysis of solved target structures. *J. Mol. Biol*., **348**, 1235–1260.
2. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.
3. 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.
4. Chothia,C. (1992) Proteins. One thousand families for the molecular biologist. *Nature*, **357**, 543–544.
5. Moult,J. (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol*., **15**, 285–289.
6. Tramontano,A. and Morea,V. (2003) Assessment of homology-based predictions in CASP5. *Proteins*, **53**(Suppl. 6), 352–368.
7. Kopp,J. and Schwede,T. (2004) Automated protein structure homology modeling: a progress report. *Pharmacogenomics*, **5**, 405–416.
8. Hillisch,A., Pineda,L.F. and Hilgenfeld,R. (2004) Utility of homology models in the drug discovery process. *Drug Discov. Today*, **9**, 659–669.
9. Peitsch,M.C. (2002) About the use of protein models. *Bioinformatics*, **18**, 934–938.
10. 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.
11. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
12. Peitsch,M.C. (1995) Protein modelling by e-mail. *Biotechnology*, **13**, 658–660.
13. Sanchez,R. and Sali,A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl Acad. Sci. USA*, **95**, 13597–13602.
14. Schwede,T., Kopp,J., Guex,N. and Peitsch,M.C. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res*., **31**, 3381–3385.
15. Koh,I.Y., Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Eswar,N., Grana,O., Pazos,F., Valencia,A., Sali,A. *et al*. (2003) EVA: evaluation of protein structure prediction servers. *Nucleic Acids Res*., **31**, 3311–3315.
16. Kopp,J. and Schwede,T. (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res*., **32**, D230–D234.
17. Pieper,U., Eswar,N., Braberg,H., Madhusudhan,M.S., Davis,F.P., Stuart,A.C., Mirkovic,N., Rossi,A., Marti-Renom,M.A., Fiser,A. *et al*. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res*., **32**, D217–D222.
18. Xie,L. and Bourne,P.E. (2005) Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comput. Biol*., **1**, e31.
19. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res*., **33**, D201–D205.
20. Melo,F. and Feytmans,E. (1998) Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol*., **277**, 1141–1152.
21. van Gunsteren,W.F., Billeter,S.R., Eising,A., Hünenberger,P.H., Krüger,P., Mark,A.E., Scott,W.R.P. and Tironi,I.G. (1996) *Biomolecular Simulations: The GROMOS96 Manual and User Guide*. VdF Hochschulverlag ETHZ, Zürich.
22. Hartshorn,M.J. (2002) AstexViewer: a visualisation aid for structure-based drug design. *J. Comput. Aided Mol. Des*., **16**, 871–881.
23. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci*., **20**, 374.
24. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci*., **23**, 358–361.
25. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*., **32**, D226–D229.
26. Pearl,F., Todd,A., Sillitoe,I., Dibley,M., Redfern,O., Lewis,T., Bennett,C., Marsden,R., Grant,A., Lee,D. *et al*. (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res*., **33**, D247–D251.
27. Velankar,S., McNeil,P., Mittard-Runte,V., Suarez,A., Barrell,D., Apweiler,R. and Henrick,K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res*., **33**, D262–D265.