Large scale protein modelling and model repository.

Manuel C. Peitsch

Geneva Biomedical Research Institute
GlaxoWellcome Research and Development
14, chemin des Aulx
1228 Plan-les-Ouates/Geneva
Switzerland
mcp13936@ggr.co.uk

Abstract

Knowledge-based molecular modelling of proteins has proven useful in many instances including the rational design of mutagenesis experiments, but it has been generally limited by the availability of expensive computer hardware and software. To overcome these limitations, we have developed the SWISS-MODEL server for automated knowledge-based protein modelling. The SWISS-MODEL server uses the Brookhaven Protein Data Bank as a source of structural information and automatically generates protein models for sequences which share significant similarities with at least one protein of known 3D-structure. We now use the software framework of the server to generate large collections of protein models. To store these models, we have established the SWISS-MODEL Repository, a new database for protein models generated by theoretical approaches. This repository is directly integrated with SWISS-PROT and other databases through the ExPASy World-Wide Web server (URL is http://www.expasy.ch).

Introduction

Insights into the three-dimensional (3D) structure of a protein are of great assistance when planning experiments aimed at the understanding of protein function and during the drug design process. The experimental elucidation of the 3D-structure of proteins is however often hampered by difficulties in obtaining sufficient protein, diffracting crystals and many other technical aspects. Therefore the number of solved 3D-structures increases only slowly compared to the rate of sequencing of novel cDNAs, and no structural information is available for the vast majority of the protein sequences registered in the SWISS-PROT database (nearly 60,000 entries in release 34)(Bairoch and Apweiler 1996). In this context it is not surprising that predictive methods have gained much interest.

Proteins from different sources and sometimes diverse biological functions can have similar sequences, and it is generally accepted that high sequence similarity is reflected in distinct structure similarity. Indeed, the relative mean square deviation (rmsd) of the alpha-carbon co-ordinates for protein cores sharing 50% residue identity is expected to be around 1Å (Chothia and Lesk 1986). Thus the most reliable prediction methods, termed comparative protein modelling (also often called modelling by homology or knowledge-based modelling), consist of the extrapolation of the structure for a new (target) sequence from the known 3D-structure of related family members (templates) (for review see Bajorath, Stenkamp and Aruffo 1993). In order to ease this process, we have established the SWISS-MODEL (Peitsch 1995) server for automated comparative protein modelling. This server is reachable through the World-Wide Web (Table I). More recently we have used the software framework of the server to build large collections of protein models, which can now be obtained from the SWISS-MODEL Repository, a new database for theoretical protein models.

Methods

Identification of Modelling Templates

Automated comparative protein modelling requires at least one sequence of known 3D-structure with significant similarity to the target sequence. In order to determine if a modelling request can be carried out, the server compares the target sequence with a database of sequences derived from the Brookhaven Protein Data Bank (PDB, Table I) (Bernstein 1977), using both FastA (Pearson and Lipman 1988) and BLAST (Altschul 1990). Sequences with a FastA score 10.0 standard deviations above the mean of the random scores and a Poisson unlikelyhood probability P(N) lower than 10^{-5} (BLAST) will be considered for the model building procedure. The choice of template structures is further restricted to those which share at least 35% residue identity with 40% of the target sequence as determined by SIM (Huang and Miller 1991).

The above procedure might allow the selection of several suitable templates for a given target sequence, and up to ten templates are used in the modelling process. The best template structure - the one with the highest sequence similarity to the target - will serve as the *reference*. All the

other selected templates will be superimposed onto it in 3D. The 3D match is carried out by superimposing corresponding Cα atom pairs selected automatically from the highest scoring local sequence alignment determined by SIM (Huang and Miller 1991). This superposition is then optimised by maximising the number of Cα pairs in the common core (Chothia and Lesk 1986) while minimising their relative mean square deviation. Each residue of the reference structure is then aligned with a residue from every other available template structure if their Cα atoms are located within 3.0 Å. This generates a structurally corrected multiple sequence alignment.

Aligning the Target and the Template Sequences

The target sequence is aligned with the template sequence or, if several templates were selected, with the structurally corrected multiple sequence alignment using the best-scoring diagonals obtained by SIM (Huang and Miller 1991). Residues which should not be used for model building, for example those located in non-conserved loops, will be ignored during the modelling process. Thus, the common core of the target protein and the loops completely defined by at least one supplied template structure will be built.

Building the Model

The co-ordinates of the model are built using *ProMod* (Peitsch 1996). The (multiple) sequence alignment serves as a correspondence table between target sequence and template structures from which a weight averaged structural framework is derived. This framework is then completed by the addition of missing or incomplete loop structures, the rebuilding of undefined backbone atoms, the correction of ill-defined side chain geometries and the addition of lacking side-chains (.Bajorath Stenkamp and Aruffo 1993; Peitsch 1996).

Model Refinement

Idealisation of bond geometry and removal of unfavourable non-bonded contacts is automatically performed by energy minimisation with CHARMM (Brooks et al 1983) using the PARAM22 parameter set and a cut-off distance for non-bonded interactions of 8 Å. The refinement of the primary model generated by *ProMod* is performed by 50 steps of steepest descent, followed by 200 steps of conjugate gradient energy minimisation.

Results and Discussion

Large Scale Protein Modelling of Protein Sequences Derived from Genome Sequencing

There is no doubt that the conclusions drawn from the scrutiny of multiple sequence alignments are far more reliable than those derived from the analysis of an isolated sequence. The mere identification of conserved and variable residues already provides useful anchor points for site-directed mutagenesis experiments aimed at the understanding of their function. Likewise, comparative structure analysis involving several members of a protein family, as opposed to single structure inspection, is expected to be invaluable for the understanding of their functional differences and for rational combinatorial library design.

Within the last two years, the full sequence of several bacterial and the yeast genomes have been determined. Thereby, ever increasing collections of truly orthologous and paralogous sequences become available. Using large scale protein modelling methods, it is possible rapidly to generate model structures for all members of a protein family. These models can then be superimposed in 3D-space and comparative structure analysis can be undertaken to understand the differences observed between species variants (for example).

A comparison of all entries of the SWISS-PROT database (rel. 34) with a non-redundant subset of all sequences with a known 3D structure (Peitsch 1997) revealed that approximately 10-15% of the known protein sequences have a suitable modelling template. For these sequences it is generally feasible to attempt knowledge-based protein modelling using a completely automated approach. This rate will of course increase as novel protein structures and folds will be solved experimentally in the coming years. However, large multi-domain complex models can presently not be built, mainly because no suitable modelling templates are available and because the prediction of protein-protein contacts is inaccurate and still in its infancy.

In order to initiate the process of large scale protein modelling, we have taken a species-based approach, and submitted - in batch mode - all known protein sequences of *Escherichia coli*, *Haemophilus influenzae*, *Mycoplasma genitalium*, *Mycobacterium tuberculosis and Bacillus subtilis* to the SWISS-MODEL server.

Model repository

All protein models generated automatically by SWISS-MODEL are annotated with information regarding the template(s) used for model building. The sequence alignment between the template and the model (target) is also provided in a numerical format. In order to ease the connectivity with other databases, each model has the same identification and accession codes as the corresponding SWISS-PROT entry. The models are stored as individual files and can be accessed through the SWISS-MODEL Web pages (Table I). In addition, every time a SWISS-PROT or a SWISS-2DPAGE entry is requested from the ExPASy Web server (Table I) (Appel, Bairoch and Hochstrasser 1994), a hyperlink to the co-ordinates file is provided if a corresponding model exists in the SWISS-MODEL Repository. Model co-ordinates can be readily downloaded and imported into the sequence to structure

Table I

A Few Relevant Internet Locations

Service	Address
ExPASy Molecular biology server	http://www.expasy.ch/
SWISS-MODEL	http://www.expasy.ch/swissmod/SWISS-MODEL.html
SWISS-MODEL Repository	http://www.expasy.ch/cgi-bin/swmodel-search-de
SWISS-PROT	http://www.expasy.ch/sprot/sprot-top.html
SWISS-2DPAGE	http://www.expasy.ch /ch2d/ch2d-top.html
SWISS-3DIMAGE	http://www.expasy.ch/sw3d/sw3d-top.html
Brookhaven Protein Data Bank	http://www.pdb.bnl.gov

workbench Swiss-PdbViewer (Guex and Peitsch 1996) or a viewing software such as RasMol (Sayle and Milner-White 1995) or Dirk Walter's Java viewer.

References

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Appel, R. D.; Bairoch, A.; and Hochstrasser, D. F. 1994. A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. *Trends Biochem. Sci.* 19:258-260.

Bairoch, A.; and Apweiler, R. 1996. The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Res.* (1996) 24, 21-25.

Bajorath, J.; Stenkamp, R.; and Aruffo, A. 1993. Knowledge-based model building of proteins: Concepts and examples. *Prot. Sci.* 2:1798-1810.

Bernstein F. C.; Koetzle T. F.; Williams G. J. B.; Meyer E. F.; Brice M. D.; Rodgers J. R.; Kennard O.; Shimanouchi T.; and Tasumi M. 1977. The protein data bank: a computer-based archival file for macromolaecular structures. *J. Mol. Biol.* 112:535-542.

Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; and Karplus, M. J. 1983. A program for macromolecular energy, minimization and dynamics calculation. *Comp. Chem.* 4:187-217.

Chothia, C.; and Lesk, A. M. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823-826.

Guex, N.; and Peitsch, M. C. 1996. Swiss-PdbViewer: A fast and easy-to-use PDB viewer for Macintosh and PC. *Protein Data Bank Quarterly Newsletter* 77:7.

Huang, X.; and Miller, M. 1991. A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.* 12:337-357.

Pearson, W. R.; and Lipman D. J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* 85:2444-2448.

Peitsch, M. C. 1995. Protein modelling by E-Mail. *Bio/Technology(Nature)* 13:658-660.

Peitsch, M. C. 1996. ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans.* 24:274-279.

Peitsch, M. C. 1997. This database was generated using all protein sequences of known 3D structure and the nrdb program of the NCBI.

Sayle. R. A.; and Milner-White E. J. 1995. RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* 20:374-376.