Protein Modeling by E-mail

From amino acid sequence to protein structure: A free one-hour service

Manuel C. Peitsch

If you submit a protein sequence to Swiss-Model, it will send you back the protein's predicted structure by e-mail. You don't have to be an expert in protein modeling, and you don't need expensive computer hardware and software. Just as you would take your roll of films to a photoprocessor, so you can now get a one-hour turnaround on converting amino-acid sequence to protein structure coordinates. The Swiss-Model service uses an automated, knowledge-based, protein modeling tool—ProMod[1,2]—implemented under the World-Wide Web (WWW; address is given in Table 1).

Swiss-Model predicts protein structure using knowledge-based, rather than de novo, modeling. Knowledge-based methods—also called "modeling by homology" or "comparative model building"—extrapolate a model for a new (target) sequence from the known 3D-structure of related family members (templates).[3] The scope for knowledge-based modeling is clear: While there are about 3000 sets of atomic coordinates in the Brookhaven Protein Data Bank (PDB),[4] the number of protein sequences in the Swiss-Prot[5] database is more than 40,000. Knowledge-based modeling works because a high sequence similarity leads to distinct structure similarity. For protein cores sharing 50% residue identity, every pair of corresponding atoms in the 2 structures are, on average, only 1 Å apart.[6]

There are drawbacks to knowledge-based protein modeling: The model's resolution is low, sometimes the results are wrong, and the method can only predict the structure of target proteins that show sequence similarities with template proteins of known, experimentally determined, structure.

But even low-resolution models can give insights into a protein's structure—insights that give experimental biology a valuable head start. Knowledge-based modeling has already proved vital in the rational design of site-directed mutagenesis experiments aimed at understanding protein function and at engineering protein properties.

Manuel C. Peitsch is a scientist at the Glaxo Institute for Molecular Biology, chemin des Aulx 14, 1228 Plan-les-Ouates, Geneva, Switzerland. (e-mail: mcp13936@ggr.co.uk).

How to Send Your Sequence

To get a protein modeled, the first step is to access the Swiss-Model server on the WWW using Internet browsing tools such as NCSA-Mosaic and

TABLE 1. Internet addresses of some protein structure-related resources and programs.

Resource/Program	Server type	Server address
Swiss-Model[a]	WWW	http://expasy.hcuge.ch/swissmod/SWISS-MODELEL.html
Swiss-3DImages[a] (17-19)	WWW	http://expasy.hcuge.ch/pub/Graphics/
		http://www.pdb.bnl.gov/images/GIF
	ftp	*host:* expasy.hcuge.ch
		directory: /pub/Graphics
		host: pdb.pdb.bnl.gov
		directory: /images
ExPASy (19)	WWW	http://expasy.hcuge.ch/
Protein Data Bank[a] (1)	WWW	http://www.pdb.bnl.gov/
RasMol[b] (11)	ftp	*host:* ftp.dcs.ed.ak.uk
		directory: /pub/rasmol
Kinemage[c] (12)	ftp	*host:* pdb.pdb.bnl.gov
		directory: /pub/kinemage
NCSA Mosaic	ftp	*host:* ftp.ncsa.uiuc.edu
		directory: /Web/Mosaic
NetScape (associated fee)	WWW	http://home.mcom.com/
	ftp	*host:* ftp.mcom.com or ftp1.netscape.com
		directory: /netscape

a These servers provide links to several resources relevant to protein structure and modeling.
b RasMol is available for UNIX workstations, Macintosh and PC-Windows.
c Kinemage, which is available for both Macintosh, and PC-Windows, is composed of Prekin to prepare the kinetic images and Mage to view them.

NetScape (Table 1). These tools are available for low-cost computers (PC-Windows and Macintosh) either free of charge or for a nominal fee.

Upon reaching Swiss-Model on the Internet, the user then completes a straightforward form, entitled "First Approach Modeling." The user types in (or cuts and pastes) either a protein sequence or its Swiss-Prot identification code, along with an e-mail address, a contact name, and a title for the request. Swiss-Model provides detailed help in Hypertext format and several links to other resources relevant to protein structure, such as the Brookhaven Protein Data Bank.

What Comes Back?

Swiss-Model returns the atomic coordinates of the final model and its 3D-profile, as well as intermediate results—such as the template proteins it proposes to use—if requested, to the user via e-mail, normally within 15 to 60 minutes.

Users can display the atomic coordinate files that come back from Swiss-Model with molecular visualization programs such as Kinemage[7] and RasMol (Table 1). These allow manipulations such as rotation, translation, and scaling of the coordinates. They can also be used to prepare color images for seminars and publications. [Another source of protein images is the publicly available collection of still (GIF) images of proteins (see "The Swiss-3DImage Collection")].

Behind the Scenes

So what at Swiss-Model between the time a user submits a protein sequence and the time he or she receives a protein structure back? The first thing Swiss-Model does is to search for template proteins, proteins of known (experimentally determined) structure that show sequence similarities to the target sequence. To do this it looks in a database (based on the Brookhaven Protein Data Bank) using BLAST,[8] FastA,[9] and SIM.[10] Although Swiss-Model will automatically select its list of template proteins, the user can define a choice of templates. If no protein of known structure shares at least 35% sequence identity with the target, modeling is not possible (and this is reported to the user by e-mail).

If sequence similarities are found, however, Swiss-Model can then produce a framework structure for the model. It does this by aligning the target sequence with the selected template sequences using a combination of sequence alignment tools and 3D-superposition. Since the optimal alignment of the template sequences is determined by superimposing their 3D-structures, it is essential that Swiss-Model ignores insertions and deletions (which are generally restricted to loops) that interrupt core secondary structure elements and that would compromise alignment. The quality of the model depends greatly on these early steps in the process: The correct identification of the conserved core of the target protein and its alignment with one or more similar template sequences. It is onto the structural framework for the target protein that nonconserved loops and side chains will be built (see below). Having constructed an averaged framework[11] from the superimposed template structures, ProMod generates its atomic coordinates.[1,2]

The subsequent steps for Swiss-Model are all, in essence, about filling in gaps in this framework structure. The first step is to rebuild the nonconserved loops from their "stems" by structural similarity searches through the Brookhaven Data Bank.[12] The next is to complete the main chain using a library of backbone elements (pentapeptides) derived from the best X-ray structures (< 2Å resolution). Then follows the addition of side chains that are not present on the template proteins, and the correction of existing side chains, using a library of allowed rotamers.[13]

Having added all the required biochemical moi-

The user types in (or cuts and pastes) either a protein sequence or its Swiss-Prot identification code, along with an e-mail address, a contact name, and a title for the request.

The Swiss-3DImage Collection

The Swiss-3DImage collection,[16,17] an Internet-based repository of more than 400 color images of 3D-structures, is set out gradually to provide views of all 3D structures in the Brookhaven Protein Data Bank (PDB). Featuring a few annotated, still images for each molecular structure, the collection endeavors to provide nonexperts with essential structural infor-mation in a comprehensive way (Figure 1). The images show key features such as disulphide bonds, bound metal ions, and residues involved in ligand-binding or enzyme activity. The images, and stereo pairs, are generated with the program RIBBONS,[18] and annotated using the IRIX Showcase (Silicon Graphics).

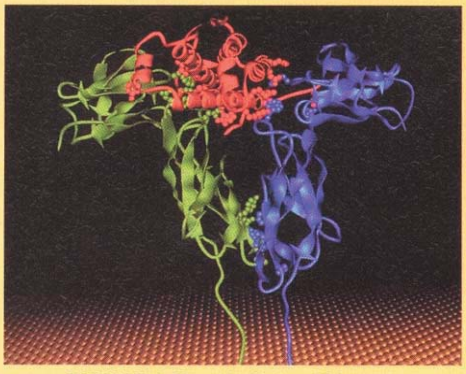

FIGURE 1. Sample Swiss-3DImages: The human tumor necrosis factor (TNF)-receptor complex.[19]

The Swiss-3DImage collection should enable scientists and students to familiarize themselves with the structural features of proteins. The images are in the public domain and can be freely downloaded from the ExPASy17 molecular biology server and the PDB. The proposed image format (.GIF) should allow anyone with an Internet connection to download and view the images using public-domain software such as Lview (PC-Windows) and Quick-GIF (Macintosh). The images are retrievable through the WWW and anonymously via *ftp* at both sites (Table 1).

Since the numbers of novel sequences and elucidated 3D-structures are increasing, more and more sequences will, probably, show homology with a protein family containing at least one member with a known 3D-structure.

eties, the model is optimized by a few cycles of energy minimization using CHARMM.[14] Model refinement, which is the last step in model building, aims to produce the best possible overall stereochemistry of the model by improving intramolecular contacts and relieving steric strain.

An assessment of the sequence-structure compatibility of the model is provided by 3D-profiles calculated using the method of Luethy *et al.*,[15] while a reliability factor, for each residue of the model, is derived from the degree of structural similarity between the templates. The visualization software displays how the reliability factor varies along the chain by color coding. Loops often have the lowest reliability.

Optimizing the Model

A weakness in the scheme, when it is operated wholly automatically (as in First Approach Modeling), is that it may produce wrong alignments early in the calculation and subsequently erroneous models. The automated alignment of moderately similar sequences is often imprecise and the boundaries of nonconserved loops are frequently ill-defined and misaligned. If a target sequence is less than 50% identical with the template, there is a 20-30% chance that the model will need to be improved.

To overcome these weaknesses, Swiss-Model's Optimize Mode will recompute a model. The user corrects sequence alignment and ProMod command files by hand, using a text editor, then accesses Swiss-Model and submits the edited sequence alignment for remodeling.

Chances of Success

Although the success rate is increasing, Swiss-Model can model less than 25% of the sequences in Swiss-Prot, and only about 10% with a high degree of reliability. But since the numbers of novel sequences and elucidated 3D-structures are increasing, more and more sequences will, probably, show homology with a protein family containing at least one member with a known 3D-structure. Hence the proportion of sequences that can be modeled should gradually increase.

References

1. Peitsch, M.C., et al. 1993. *Int. Immunol.* **5:**233-238.
2. Peitsch, M.C., et al. 1995. *Mol. Immunol.* In press.
3. Bajorath, J., et al. 1993. *Protein Science* **2:**1798-1810.
4. Bernstein, F.C., et al. 1977. *J. Mol. Biol.* **112:**535-542.
5. Bairoch, A., et al. 1992. *Nucleic Acids Res.* **20:**2019-2022.
6. Chothia, C., et al. 1986. *EMBO J.* **5:**823-826.
7. Richardson, D.C., et al. 1992. *Protein Science* **1:**3-9.
8. Altschul, S.F., et al. 1990. *J. Mol. Biol.* **215:**403-410.
9. Pearson, W.R., et al. 1988. *Proc. Natl. Acad. Sci. U.S.A.* **85:**2444-2448.
10. Huang, X., et al. 1991. *Adv. Appl. Math.* **12:**337-357.
11. Blundell, T., et al. 1987. *Nature* **326:**347-352.
12. Greer, J. 1990. *Proteins* **7:**317-334
13. Ponder, J.W., et al. 1987. *J. Mol. Biol.* **193:**775-.
14. Brooks, B.R., et al. 1983. *J. Comp. Chem.* **4:**187-217.
15. Luethy, R., et al. 1992. *Nature* **356:**83-85.
16. Peitsch, M.C., et al. 1995. *Trends Biochem. Sci.* **20:**82-84
17. Appel, R.D., et al. 1994. *Trends Biochem. Sci.* **19:**258-260.
18. Carson, M. 1991. RIBBONS 2.0 *J. Appl. Cryst.* **24:**958-961.
19. Banner, D. W., et al. 1993. *Cell* **73:**431-445.