Internet On-Ramp

Molecular Modeling on the Web

**Shoba Ranganathan, eBioinformatics,
Sydney, Australia (Shoba.Ranganathan
@eBioinformatics.com)**

With the worldwide genomic sequencing efforts producing a steady stream of data, the sequences of tens of thousands of proteins can now be determined. Understanding protein function requires knowledge of protein structure. Information on the 3-D structure of a protein can provide insight to the protein's interactions with other molecules; this information can also be exploited in such endeavors as protein engineering, mutagenesis, and computer-aided drug design. Detailed structural information is available for only 10% of all known protein sequences. Of the remaining 90%, a homologous protein of known structure has been identified for 10%–25%. For this subset, a rapid *in silico* approach to 3-D structure is accomplished using homology or comparative modeling. For another 25% of the available protein sequences, structural models can be generated using secondary and tertiary structure prediction methods.

Until recently, many biologists have been reluctant to use molecular modeling as a computational tool because of the high costs of modeling software, its user “unfriendliness”, and the expensive hardware (usually UNIX® platforms) dictated by the software. Biologists have thus been hindered from making critical associations relating biological structure to function. So, if you have a protein sequence (referred to as the query sequence) and would like to build a molecular model using Internet-based tools, here is a route map.

Understanding Protein Structure

An understanding of protein structure is a prerequisite to molecular modeling. The Protein Databank (PDB; www.rcsb.org/pdb) is the single worldwide archive of structural data on biological macromolecules, primarily proteins. The structural information in the PDB is derived primarily from experimental studies using electron microscopy, X-ray crystallography, and NMR. On the structural scale, a protein can span four levels: primary, secondary, tertiary, and quaternary structures. A brief introduction to these and similar terms used in describing protein structure is available on the Web (<http://molmod.angis.org.au/~shoba/asbmb/>). A more detailed introduction to proteins and structures can be accessed at the PDB site itself (<http://www.rcsb.org/pdb/education.html>). A tip for first-time visitors to the PDB is to start with a structure itself and follow all the Web pages describing that structure. My favorite is 1SHA (the PDB identifier or ID code), an example of the phosphotyrosine recognition domain. Simply enter “1SHA” in the box next to “Enter a PDB ID” under the

“Search” heading at the left of the PDB Web page.

When you do not know the PDB ID, you can search the PDB using “SearchLite” or SearchFields” and look for structures containing the keywords that you have specified. The PDB database entry itself is a very large text file that contains a descriptive section called the header and the x, y, and z coordinates for each atom of the macromolecule. Normally, only the “heavy” atoms are listed, excluding the hydrogen atoms. It is nearly impossible to read this text file and figure out how the protein is folded without the aid of graphic viewers such as Rasmol (www.bernstein-plus-sons.com/software/rasmol/) or the Web browser plug-in, Chime (<http://www.mdli.com/support/chime/default.html>).

Is There a Structural Homologue?

Identifying a structural homologue simplifies the molecular modeling problem to one of homology model building. BLAST and PSI-BLAST searches of a new protein sequence against the sequences in the PDB database can be used to locate structural templates. One of the problems resulting from database matching is figuring out which of the hits to base the model on. A match spanning the entire length of a sequence and with a good score can be considered good. A poor score over a small match length (less than 25% of the query sequence) is not considered significant, and secondary or tertiary prediction methods will then have to be tried.

The homology modeling approach works well when the query protein has close homologues of known structure and is at the basis of the SWISS-MODEL (2) automated structure prediction system (<http://www.expasy.ch/swissmod/SWISS-MODEL.html>). The Prodom database of protein domains (<http://www.toulouse.inra.fr/prodom.html>) provides a BLAST server to search for domains contained in a query sequence and allows this information to be passed directly to the SWISS-MODEL server for subsequent protein structure modeling. Another completely automated service for homology models is the CPH Model server (<http://www.cbs.dtu.dk/services/CPHmodels/>), which uses a FASTA search to locate homologues within the PDB and then performs rapid homology modeling using distance constraints predicted by neural networks and probability density functions (5). The procedure works well for single domain sequences that are highly homologous to a structure but fails for multidomain protein sequences. For a quick look (the “rough and ready” approach), this server will generate a homology model complete with coordinates in the PDB format.

The integrated Web-based bioinformatics analysis system BioNavigator (<http://www.bionavigator.com>) takes a more “manual” approach to structure prediction using the same principle: the researcher can run a BLAST search of the PDB protein structure database and then enter the query sequence and its matching

Internet On-Ramp

structure into other programs for building a structural model of the query protein based on the structure. The latter method gives the researcher complete control over the entire model-building process.

Proteins often contain several modules or domains, each with a distinct evolutionary origin and function. The new CD-Search service from NCBI (<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>), based on the reverse position-specific BLAST algorithm, is very useful in identifying the conserved domains present in a protein sequence. Since these conserved domains correspond to compact structural units, it is possible to identify individual domains of multidomain proteins. If you do not find any significant hits with the default parameters, you might be able to locate more remote homologues by following the link for "advanced options" and increasing the expectation value.

Template Selection

When there is more than one structural homologue available, the template for building the model will have to be selected carefully, in accordance with rules for comparative protein modeling (7). Generally, an X-ray structure is preferred to an NMR structure, and among X-ray structures, the one with the lowest resolution is chosen. Normally, more than one template is used only when the extra templates contribute to significant new structural information.

Lining Up the Query and the Template Sequence

Model-building techniques depend on the alignment of the query sequence with those of the templates. The matched blocks in the alignment usually represent the structurally conserved regions (SCRs) and include secondary structural elements, while the insertion-deletion segments (indels) in the alignment represent the structurally variable regions (VRs), usually loops. Alignments are usually improved by using a set of homologous sequences, and the many programs available for this step have been reviewed recently (3).

Building the Model

Based on the alignment, the SCRs of the query sequence are constructed by simply copying the backbone and building new side-chains. The VRs are then modeled by joining the appropriate backbone atoms for deletion regions or by building in new segments for the insertion regions, normally using loop generation methods (4).

Besides automated model building available from the SWISS-MODEL server, the WHAT IF server (8) (www.cmbi.kun.nl:1100/WIWWWI) can generate a homology model based on the researcher's alignment with a single template structure. However, WHAT IF does not handle VRs in the alignment, resulting in structural models with chain breaks. Nevertheless, the WHAT IF

structural model is very useful especially for side-chain conformations and can be used for *in silico* experiments such as mutations and disulfide bridge location before wet-lab experimentation.

A high-quality structural model with all VRs and loop refinements using Monte Carlo simulations can be built using the ICM (1) methodology via the BioNavigator interface. The input required here are the query sequence and an alignment with a single structural template.

How Good Is the Model Structure?

The quality of the model is estimated using structural quality analysis tools. Complete analysis can be done using WHAT CHECK (6), which is available as part of the WHAT IF package from the WHAT IF server. Ramachandran plots for checking the backbone torsions can be generated using the ICM package described above.

Beyond Molecular Modeling

The structural model is only the beginning of a series of *in silico* experiments that can be performed. With a model, it is possible to look for functional residues at an active site, based on the location of key residues, which can then be mutated in the laboratory. Stabilizing a protein by incorporation of disulfide bridges is a common technique used in molecular biology and biochemistry. Instead of mutating pairs of residues at random, the structural model can be used to scan for the best location of a disulfide bridge, serving to direct experimental research. Such computer-based experiments are easily accessible through the previously mentioned WHAT IF and BioNavigator Web sites.

The author's current address is Shoba Ranganathan, Bioinformatics Centre, National University of Singapore, Singapore. e-mail: shoba@bic.nus.edu.sg.

REFERENCES

1. Cardozo, T., M.M. Totrov, and R.A. Abagyan. 1995. Homology modeling by the ICM method. *Proteins Struct. Funct. Genet.* 23:403-414.
2. Guex, N., A. Diemand, and M.C. Peitsch. 1999. Protein modeling for all. *Trends Biochem. Sci.* 24:364-367.
3. Gaskell, G.J. 2000. Multiple sequence alignment tools on the Web. *BioTechniques* 29:60-62.
4. Johnson, M.S., N. Srinivasan, R. Sowdhamini, and T.L. Blundell. 1994. Knowledge-based protein modeling. *Crit. Rev. Biochem. Mol. Biol.* 29:1-68.
5. Lund, O., K. Frimand, J. Gorodkin, H. Bohr, J. Bohr, J. Hansen, and S. Brunak. 1997. Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.* 10:1241-1248.
6. Rodriguez, R., G. Chinea, N. Lopez, T. Pons, and G. Vriend. 1998. Homology modelling, model and software evaluation: three related resources. *Bioinformatics* 14:523-528.
7. Sali, A. and J.P. Overington. 1994. Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.* 3:1592-1596.
8. Vriend, G. 1990. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* 8:52-56.