Web alert

Molecular modelling

Chemistry & Biology March 1999, **6**:R83-R84

 ISSN 1074-5521

It is a truism to note that molecular modelling means different things to different people — a broad working definition is that it includes the calculation of two- and three-dimensional structures, as well as calculation and display of molecular properties. The internet is useful not only for finding out about the latest developments, but increasingly for actually making models (e.g. Swiss-Model, see below). It also means that sophisticated programs can be made available to scientists who do not have extensive training or expensive graphics equipment (e.g. GRASS, http://trantor.bioc.columbia.edu/GRASS/surfserv_enter.cgi, for visualisation of molecular surfaces).

Three different viewpoints about what molecular modelling is can be found at the MathMol site (http://www.nyu.edu/pages/mathmol/quick_tour.html); in a lecture course given by Henry Rzepa at the Department of Chemistry, Imperial College (http://www.ch.ic.ac.uk/local/organic/mod/); and in the variety of modelling tutorials listed by the National Institutes of Health (NIH) Center for Molecular Modeling (http://cmm.info.nih.gov/modeling/education.html). Two recent online reviews by Christopher Smith, published in *HMS Beagle*, are very useful for protein modellers: Molecular Modeling, an Internet Resource for Biologists (http://www.biomednet.com/hmsbeagle/41/webres/insitu.htm) describes structure, sequence and visualisation resources, and the second (http://www.biomednet.com/hmsbeagle/41/webres/wreview.htm) focuses on metalloproteins.

Many different programs are used by modellers, to answer questions ranging from 'how many possible isomers of a given molecular formula are there?' (see MOLGEN at http://www.mathe2.uni-bayreuth.de/molgen4/) to 'how does one go about docking a small molecule into a protein's active site?' (e.g. http://www.cmpharm.ucsf.edu/kuntz/dock_demo.html). Standard operations include, for example, retrieval of molecules from databases, construction of molecules, as well as conversion from two to three dimensions, geometric operations on coordinates, protein sidechain substitution, energy minimisation, molecular dynamics, calculation of hydrophobic or electrostatic properties and various forms of display.

Comprehensive lists of programs for these operations are available at, for example, the CHEMIE.DE site (http://www.chemie.de/), which describes and reviews each program (including availability and cost). For protein modelling, there are extensive listings of interactive databases and software packages at ExPASy (http://expasy.hcuge.ch/www/expasy-top.html), run by the Swiss Institute of Bioinformatics, and at Network Science (http://www.netsci.org/Resources/Software/top.html). There you will find the tools required to go from the gene sequence to the secondary structure and, in favourable cases, to a three-dimensional model (searching for similarities with other known proteins or structures is also possible).

Three-dimensional structures of biological macromolecules determined by X-ray crystallography and NMR spectroscopy are well-organised in online databases such as the Protein Data Bank (PDB; http://www.pdb.bnl.gov, currently undergoing a transition to the Research Collaboratory for Structural Bioinformatics, RCSB; http://www.rcsb.org/) and the Nucleic Acid Database (http://ndbserver.rutgers.edu/NDB/). Each database has extensive search and retrieval facilities.

It is also worth investigating if your favourite protein has already been modelled. Although it is still not possible to make models *ab initio* from the amino-acid sequence alone, comparative protein modelling is being applied to sequence databases, to generate models for sequences predicted to have the same fold as a related structure deposited in the PDB. ModBase, from Andre Sali's Lab at Rockefeller University contains approximately 15,000 reliable models for substantial segments of ~4000 proteins in the *Saccharomyces cerevisiae*, *Mycoplasma genitalium*, *Methanococcus jannaschii*, *Caenorhabditis elegans* and *Escherichia coli* genomes [1].

3D Crunch, a large-scale protein modelling project, was run in 1998 (http://www.expasy.ch/swissmod/SM_3DCrunch.html) as a collaboration between Silicon Graphics, who provided a CRAY Origin2000 server, and scientists from the Swiss Institute of Bioinformatics, GlaxoWellcome, the Imperial Cancer Research Fund and the Lyon Bioinformatics Centre. The aim was to predict the structures of as many proteins as possible in the SWISS-PROT and TrEMBL databases; the results are available in searchable form on the web. The Swiss-Model server is an automated protein modelling server running at the GlaxoWellcome Experimental Research in Geneva, Switzerland (http://www.expasy.ch/swissmod/SWISS-MODEL.html). Again this requires a related protein to have already been solved, and as it runs with no human intervention, careful inspection of the results is recommended.

There are also many databases derived from the PDB (see *Chem. Biol.* **5**:R149-R150) but two that span the small-molecule/macromolecule interface are HIC-UP, the Hetero compound Information Centre at Uppsala (http://alpha2.bmc.uu.se/hicup/), a resource for structural biologists dealing with hetero compounds ('small molecules') encountered in files from the PDB and RELIBase (http://www2.ebi.ac.uk:8081/home.html), an archive for structural data about receptor–ligand

complexes. The main purpose of RELIBase is to provide selective and efficient access to the receptor–ligand complexes currently deposited in the PDB and to make the enormous wealth of information contained in the receptor–ligand structures available for structure-based drug-design studies.

Information about small molecules is less centralised. The Inorganic Crystal Structure Database (http://www.fiz-karlsruhe.de/fiz/products/icsd_.html) and the Cambridge Structural Database (CSD) at the Cambridge Crystallographic Data Centre (http://www.ccdc.cam.ac.uk) are definitive databases of both inorganic and organic crystal structures, but neither the data nor the search tools are freely available on the web. Compilations of structures are available at several sites (in any one of a number of formats) with varying amounts of chemical information. Beware the confusion that can arise when a site advertises 'molecule X is available in .gif and .pdb format'. The .gif may be both a useful and beautiful picture, but only the .pdb format contains the three-dimensional coordinates required for modelling. As a starting point, the RasMol homepage (http://klaatu.oit.umass.edu:80/microbio/rasmol/) indicates several sources of three-dimensional coordinate data (in .pdb format), in addition to providing two excellent free viewers, RasMol and Chime. The NCI database browser provides access to nearly 250,000 structures, and is searchable, for example, by chemical abstract service (CAS) number, elemental formula or (sub)structure expressed either as a SMILES string or drawn in a java applet. Output can be returned in a variety of formats, including three-dimensional coordinates (when available).

The MathMol library (http://www.nyu.edu/pages/mathmol/library/Overview.html) contains three-dimensional structures for assorted molecules, from water to macromolecules, discussed in introductory biology and chemistry textbooks. Klotho (http://www.ibc.wustl.edu/klotho/) contains data and coordinates for biochemical compounds; the Chemistry Department at Okanagan University College offer over 1100 compounds (http://www.sci.ouc.bc.ca/chem/molecule/molecule.html); and Robert Lancashire, at the University of the West Indies, maintains a database that includes both transition-metal complexes and organic molecules in .pdb format (http://wwwchem.uwimona.edu.jm:1104/spectra/pdbIndex.html). More specialised collections that include three-dimensional structures together with other relevant data include the Smells Database (http://mc2.cchem.berkeley.edu/Smells/index.html) and the structures from the published technical reports at the National Toxicology Program of the National Institutes of Health (http://ntp-db.niehs.nih.gov/Main_Pages/pub-Structures.html) are in .mol format. CambridgeSoft provides a database of over 75,000 'micromolecules' that can be searched at no cost with the utility ChemFinder (http://chemfinder.camsoft.com). The resulting structure files are in a proprietary format that can be displayed by ChemDraw or Chem3D, but not the usual free viewers, such as RasMol. If, however, you convert the structure to SMILES format (http://www.daylight.com/dayhtml/smiles/) it can be submitted to CORINA, which can then return a .pdb file. CORINA is a rule- and data-based system that automatically generates three-dimensional atomic coordinates. Its web interface (http://www2.ccc.uni-erlangen.de/services/3d.html) allows queries to be presented in several formats.

Using the programs described above will quickly demonstrate that neither the program systems nor the data formats are easily interoperable. Various integrated systems are being developed and used, however, both as free tools on the web and as intranet applications. One example is the Molecular Modelling Toolkit (http://starship.skyport.net/crew/hinsen/mmtk.html), a program library for molecular-modelling applications, that programmers can adapt to their own problems. Although intranet applications are not distributed, it is useful to read articles such as that by Peter Ertl and Olivier Jacob (A WWW-based Chemical Information System; http://www.elsevier.com/homepage/saa/eccc3/paper6/) presented at the Third Electronic Computational Chemistry Conference, to show how web-based tools are used within a large company. An article by Omer Casher *et al.* (http://www.ch.ic.ac.uk/rzepa/RSC/P2/4_05970K.html) explains some of the technicalitites of handling molecular information on the web. Further developments in markup language beyond HTML to XML (eXtensible Markup Language) facilitate the development of other types of markup in structured documents, such as Chemical Markup Language (CML; http://www.xml-cml.org/). The goals of CML include platform and application interoperability (avoiding awkward file format conversions), support for complex documents and electronic publication, as well as interaction with other XML metadata standards, such as MathML (see http://www.w3.org/Press/1998/MathML-REC.html).

Modelling technology is constantly developing and happily the internet seems set to continue to provide not only information about the latest developments, but an expanding variety of servers to assist in the process.

References

1. Sali, A. (1998). 100,000 protein structures for the biologist. *Nat. Struct. Biol.* **5**, 1029-1032.

Judith Murray-Rust, Crystallography Department, Birkbeck College, London, WC1E 7HX, UK; ubcg09j@mail.cryst.bbk.ac.uk.
Peter Murray-Rust, Department of Pharmaceutical Sciences, Nottingham University, Nottingham, NG7 2RD, UK; Peter.Murray-Rust@nottingham.ac.uk