Internet On-Ramp

Molecular Biology Databases on the Internet, Part II

David M. Sander, Tulane University
School of Medicine, New Orleans, LA, USA
dmsander@mailhost.tcs.tulane.edu

Databases containing information of interest to molecular biologists continue to prosper on the Internet. In a previous **Internet On-Ramp** (September 1996), a general introduction was given on molecular biology databases found on the World Wide Web (WWW). That article is still available—either on your bookshelf with all of your other valuable *Biotechniques* issues or on the WWW (**http://www.tulane.edu/~dmsander/biotechniquessites.html**). In this second review, I would like to demonstrate how specific data elements of the larger databases (EMBL, GenBank®, SWISS-PROT) are being rearranged into targeted research tools with significant power. With a focus on sequence analysis, motif identification and other data such as enzyme specificity and multiple sequence analysis, these tools, rather than being made commercially available, are being developed by a variety of database providers and made available on the WWW.

Simplified Access to Multiple Databases

Means of accessing multiple databases using one central utility are increasingly available and becoming more user friendly. An example of this trend is DBGET (**http://www.genome.ad.jp/dbget/dbget.links.html/**). After accessing this Uniform Resource Locator (URL) using an Internet browser (Netscape Navigator, Internet Explorer, etc.), the user is presented with a graphic display illustrating a wide variety of databases. By clicking on portions of the figure, searches can be made of these databases using a unified syntax. Rather than directly querying dozens of different databases on the Internet, you can use this database retrieval system to access them individually using this common interface. DBGET currently supports access to more than sixteen databases including GenBank and European Molecular Biology Laboratory (EMBL) for nucleic acid sequences; SWISS-PROT, Protein Information Resource (PIR®), Protein Research Foundation (PRF) and PDBSTR for protein sequences; Brookhaven Protein Data Bank (PDB) for 3-D molecular structures; PROSITE, EPD and TRANSFAC to identify sequence motifs; LIGAND for enzyme reactions; PATHWAY for metabolic pathways; PMD for amino acid mutations; Online Mendelian Inheritance in Man (OMIM) as an index to genetic diseases; among others. While some of these individual databases have been described previously, others will be introduced below. Similar services to DBGET are provided by EMBL's SRS (**http://www.embl-heidelberg.de/srs/srsc**) and National Center for Biotechnology Information (NCBI's) Entrez (**http://www3.ncbi.nlm.nih.gov/Entrez/index.html**). Resources such as these serve to simplify the complexity faced by a new Internet user in search of data and therefore serve as a great place to begin your search.

Protein Sequences

Protein sequence analysis is generally performed to either obtain an accurate alignment of a novel sequence with known proteins or to determine aspects of a protein's structure by comparison with known structural elements. Both of these aims can be accomplished though the Internet and are described by an on-line tutorial for protein sequence analysis at the University of Oxford. "Protein Sequence Alignment and Database Scanning" (**http://geoff.biop.ox.ac.uk/papers/rev93_1/rev93_1.html**) details the considerations necessary to obtain accurate data through Internet sources. This site goes into some detail with topics including: database

news.group.news

News.group.news goes to cyberspace and back to gather Net news for you. We emphasize practical, methods-related issues that are discussed in many newsgroups and also feature issues of general interest to biological scientists. But be warned! On this page of *BioTechniques*, we don't *review* it — we just *report* it!

Seeking a **protein determination method** that could be applied to cell lysates in SDS-PAGE sample buffer, a poster on **bionet.molbio.methds-reagnts** was advised to consider the approach described in Winterbourne, D.J., 1993, Chemical assays for proteins, Methods Mol. Biol. *19*:197-202; a procedure that involves precipitating the protein on filter paper, leaving most interfering substances in solution.

Commenting on the use of **RAPDs for phylogenetic analysis**, one expert-on-line at **bionet.molbio.rapd** recommends maximizing the number of polymorphic bands you look at. To determine whether your results are consistent, analyze the data after collecting 50, 100 and 150 polymorphic bands.

Too much salt in your oligos? Netters on **bionet.molbio.methds-reagnts** strongly recommend the procedure for desalting with minimum loss of DNA described in B. Freie and S.H. Larsen, 1991, Oligonucleotide purification in milligram quantities, BioTechniques *10*:420-422.

Seeking information on **microsatellites in *Arabidopsis,*** a Netter on **bionet.genome. arabidopsis** was advised to search the "papers" section of the prototype *Arabidopsis thaliana* database, accessible through the *At*DB home page at **http://genome-www.stanford.edu/Arabidopsis**.

Concerned that a dramatic decrease in yield of PCR product accompanied by the appearance of artifactual bands may be due to **freeze-thaw degradation of primers**, a poster was advised to store primer stocks at -20°C, but to maintain working solutions at 4°C.

Current and past postings of all the **BIOSCI/bionet newsgroups** are accessible on the BIOSCI Web site at **http://www.bio.net.**

BioTechniques Editorial Department

scanning, the comparison of two sequences, amino acid scoring schemes, multiple sequence alignment and assessing alignment accuracy.

Other sites of interest for protein sequence analysis include Prot-Web (**http://brut.gdb.org/**)—a collection of databases that offer three primary protein database search utilities. The first is the PIR (**http://www.gdb.org/Dan/proteins/pir.html**), which searches several databases using different protein parameters including keywords, journal references, molecular weight and motif. The OWL database (**http://www.gdb.org/Dan/proteins/owl.html**) is designed as a nonredundant protein sequence database with entries from SWISS-PROT, PIR, GenBank translations and NRL-3D. Some of the entries found using OWL will have images of the proteins queried. The NRL-3D (**http://www.gdb.org/Dan/proteins/nrl3d.html**) database, which focuses on sequence/structure relationships, yields search results with protein images and a plethora of three-dimensional structural information. Good tutorials located at Prot-Web make the site very user friendly.

Protein Motifs

Another aspect of protein sequence analysis is the identification of protein motifs. One such utility exists at the Hutchinson Cancer Research Center. Termed BLOCKS (**http://www.blocks.fhcrc.org/**), this program utilizes multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins as aids to detection and verification of protein sequence homology.

A site offering similar features is PRINTS, the Protein Motif Fingerprint Database (**http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html**). PRINTS collects protein fingerprints or groups of conserved motifs used to identify protein families. These fingerprints also account for the folding of proteins; therefore, PRINTS adds more flexibly and power into searches than could be achieved using single motifs. The fingerprints have been developed using the OWL database and are available for WWW-based interaction.

DNA and RNA Motifs

In a similar line of thought, the need for specific sequence motif identification in RNA and DNA has spawned several databases. One is called TRANSFAC (**http://transfac.gbf-braunschweig.de/TRANSFAC/index.html**). It compiles data from other databases about gene sequences that have a role in transcriptional regulation and builds programs for the identification of potential promoter or enhancer sites. Even a casual user can generate results of interest.

For example, on the search page (**http://transfac.gbf-braunschweig.de/cgi-bin/QueryTransfac/search.pl**), I entered “pit-1” in the open box and selected a search for “factor” by use of the “sites table” and “list of links”. This produced information including a transcription factor classification for pit-1, list of relevant species, links to EMBL and SWISS-PROT for raw pit-1 sequences, a bibliography and additional details. Other options at the site include a full classification system for transcription factors, the ability to browse the thousands of entries by a number of parameters and on-line documentation.

Other DNA/RNA motifs that molecular biologists are commonly seeking include restriction endonuclease sites. REBASE (**http://www.gdb.org/Dan/rebase/rebase.html**) integrates a large variety of information about each restriction enzyme or methylase into a single report including: organism of origin, recognition sequences, methylation specificity and commercial availability. A similar on-line enzyme nomenclature database called ENZYME (**http://expasy.hcuge.ch/sprot/enzyme.html**) collates information about the nomenclature of enzymes. Based on the findings of the International Union of Biochemistry and Molecular Biology (IUBMB), it contains information about each type of characterized enzyme.

Multiple Sequence Alignments

Once only available locally, Internet browser-based forms for performing a multiple sequence alignment are now available on the WWW. A good example is the MSA site (**http://alfredo.wustl.edu/msa.html**) at Washington University. This site allows the user to input as many as eight protein sequences for multiple alignment. However, rather than entering them directly, you have the option of specifying accession numbers from SWISS-PROT or PIR. Results are returned in the form of a Web page or through e-mail. Various parameters are adjustable within the algorithm. The Multalin Multiple Alignment (**http://www.ibcp.fr/multalin.html**) utility at the IBCP in France can perform similar functions but gives results only by e-mail.

PCR Primers

As a final example of the utility of sequence-based databases, a PCR primers database is now available (**http://www.ebi.ac.uk/primers_home.html**), which attempts to index primers used in basic research while excluding primers associated with megabase sequencing projects. By limiting its resources to these selected functional primer sets, it has the potential of yielding significant temporal and financial savings to molecular biologists world wide. The database is accessed through a variety of interfaces, including direct primer sequence submission, target, sequence, species, contributing author, etc. using a forms-based query site (**http://www-srs.caos.kun.nl/srs/srsc**) within EMBL’s SRS multiple database accession utility.

Conclusions

While many readers may be alternately enthused or bored by the WWW sites that I’ve chosen to review, this article and its predecessor only scratched the surface of biological databases and information available on the Internet. With a little bit of searching using index pages like WWW-Virtual Library Biomolecules section (**http://golgi.harvard.edu/sequences.html**), the molecular biology research tools at the CMS Molecular Biology resource page (**http://www.unl.edu/stc-95/ResTools/cmshpa.html**) or those listed by the Pasteur Institute (**http://www.pasteur.fr/other/biology/english/bio-databases-uk.html**), you will find a wide variety of other data sources including databases focusing on 2-D protein gel analysis, 3-D molecular structures, metabolic pathways and phylogenies. For additional education, the EMBnet Biocomputing Tutorials (**http://www.hgmp.mrc.ac.uk/Embnetut/Universl/embnettu.html**) serve as a great intermediate step in learning about genomic databases on the Internet. With continuing increases in both the specificity and power of these utilities and the sophistication of Internet biologists, the future of molecular biology on the Internet is very promising.

The text of this article will be available with its assorted links at: **http://www.tulane.edu/~dmsander/biotechniquessites.html**.