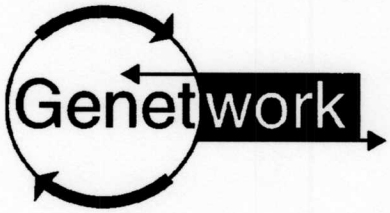

Network sequence retrieval

Box 1. Dedicated network sequence retrieval programs

Users who spend a significant amount of time working with sequences will probably want a more sophisticated or convenient means of accessing data. The leader in this field is Network Entrez (Nentrez) (http://www.ncbi.nlm.nih.gov/Search/client.html), which provides the features of the WWW version, but is more interactive. The program runs on X Windows, and Macintoshes. The source code to Entrez is included in the NCBI Toolbox routines (ftp://ncbi.nlm.nih.gov/toolbox) which are all freely available for those who wish to do any sort of sequence access and manipulation. The library was used to create nclever (gopher://megasun.bch.umontreal.ca/00/CMB/Entrez/Clever/About), that provides many features of Entrez for users without a graphical interface on their networked computer; it is ideal for researchers connected to a Unix machine by a serial line or a modem. DNA Workbench, (ftp://cbil.humgen.upenn.edu/) constructed by James Tidsall at the University of Pennsylvania, is a Perl-language program that runs on Unix, Macintoshes and PCs and allows users to create their own Perl routines to select and manipulate sequences in the GenBank and PIR databases.

Retrieving DNA and protein sequences from a database is one of the common computer tasks for molecular biologists and should be one of the simplest. However, despite (or perhaps because of) the proliferation of systems for storing sequences and finding them by the author's name, a description of the sequence, or an accession number, even this apparently trivial task can be daunting. The retrieval programs themselves are generally well designed and fairly easy to use, especially if you have the manual sitting on your lap. However, they require that you not only have the program set up and installed on your machine, but also that you have an up-to-date version of a large database, in an appropriate format. That's a lot of trouble to go through if, for example, you just want the sequence of insulin.

Customized network clients, such as the excellent Nentrez program by the National Center for Biotechnology Information (NCBI), save the hassle of maintaining the sequence databases. Instead, these programs connect over the network to a central database which is kept up to date by the maintainers. Indeed, soon GenBank will be so large that it will probably stop being distributed on CD-ROMs. But for scientists who wish to spend their research time at the bench and not at the computer, even the trouble of obtaining current versions of the software, installing them and learning about them can be a distressingly large time investment.

A World Wide Web (WWW) client can provide a one-piece solution. You still need to obtain a program, the browser, but generally this is much less of a task than setting up any other sequence access system. To gain market share of the huge number of users on the Web, several companies have developed slick and easy-to-use viewers. While the eventual goal will be to separate users from their money, for now all popular viewers are available free of charge to academic researchers. The Web client isn't limited just to a single purpose, so once you've set it up and learned how to use it, you will be able to access an enormous number of other scientific (and non-scientific) resources. Time spent learning how to use the WWW is a good investment.

Organizations around the world have supplied the Web with a cornucopia of different sequence retrieval systems (see Boxes 1, 2). The WWW interface for Entrez (http://atlas.nlm.nih.gov:5700/ Entrez/index.html) is neither the easiest nor the savviest system, but provides a good balance of reliability, comprehensiveness and useful features. Originally developed in 1991 for use with a CD-ROM database, Entrez was prescient in its use of links within and between its four component databases (proteins, nucleotides, Medline, and taxonomy). Two flavours of Web interfaces are provided, one intended for modern WWW viewers with support for 'forms', and one for the older and more primitive browsers.

While the form interface has more options, the latter version is very simple to use. To find a protein's sequence (1) go to the Entrez home page, (2) choose 'search protein database' with non-forms interface then (3) choose 'search by text terms' and (4) enter a term such as 'cro'. Rather than immediately showing the sequence, the system brings up a dizzying array

Box 2. Selected WWW sequence retrieval sites

The Sequence Retrieval System (SRS) (http://www.ebi.ac.uk/srs/srsc) does not collect biological data of its own but, by integrating some two dozen networked databases, it is the uncontested champion of biological information interconnection. Unfortunately, it is currently rather difficult to use. SWISS-PROT, at the University of Geneva (http://www.expasy.ch/), was one of the first major databases to provide hypertext access. It presents normal-looking entries, with a wide variety of cross-references available as hyperlinks. Other useful sites for sequence retrieval are at the EBI (http://www.ebi.ac.uk/), NCGR (http://www.ncgr.org/), NCBI (http://www.ncbi.nlm.nih.gov/), GenomeNet (http://www.genome.ad.jp) and Harvard (http://golgi.harvard.edu/sequences.html).

Genetwork is a regular column of news and information about Internet resources for researchers in genetics and development. Readers wishing to make use of the systems described here will need to have machines configured to access networks and have standard software available. Genetwork is compiled and edited with the help of Steven E. Brenner (MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK CB2 2QH) and Jeremy Rashbass (Department of Histopathology, Addenbrooke's Hospital, Hills Road, Cambridge, UK CB2 2QQ).
If you would like to announce or publicize an Internet resource, please contact: TIG@elsevier.co.uk.

of new choices. Selecting the report format for RCR_BP434 brings up, unsurprisingly, a report on the sequence of the bacteriophage 434 Cro repressor protein. The FASTA and ASN.1 options provide the sequence in formats more frequently used by programs.

The additional options available are links to other entries in the Entrez system. The Medline link retrieves information (with abstract available) about the article in which the sequence of the protein is published. The next hyperlink, 'Protein neighbors', means that there are other proteins in the database which have significant sequence similarity to 434 Cro. These have been identified by running a BLAST (Ref. 1) search of the entire protein database against itself. Selecting that hyperlink would display the list of these similar proteins. Finally, choosing the nucleotide link would find the DNA sequence of phage 434 that contains the *cro* gene.

Scrolling a little further down the page obtained for 'cro' yields a strange result: other entries also contain information and the sequence of the very same protein as RCRO_BP434. This is because Entrez is an integrated database, combining sequences from many sources: proteins come from SWISS-PROT, PIR, the Protein Data Bank and translations of coding sequences in GenBank. Usually, identical sequences are merged into a single entry, but occasionally a gremlin gets loose and identical twins aren't reunited.

Neighbours in the different component databases mean different things. As mentioned above, protein neighbours are similar, and thus probably homologous, proteins. In the DNA database, neighbours are sequences which, based on a BLAST comparison, appear to overlap the selected sequence; thus, neighbours can be used to find increasingly large regions of a genome's sequence. Medline neighbours are found by a process analogous to a homology search of title and abstract text; neighbours are similar articles. Only a subset of Medline is available through Entrez, but the neighbouring feature makes it not only a cheap way of doing literature searches, but also one of the most powerful.

Perhaps the greatest limitation of WWW Entrez is that, because it is also distributed as a standalone database, it has few links to external databases. However, NCBI is now devoting more effort to this interface, so more connections to other resources are probably on the way.

Steven E. Brenner
S.E.Brenner@bioc.cam.ac.uk

Reference

1 Altschul, S.F. *et al.* (1990) *J. Mol. Biol.* 215, 403–410

HUM-MOLGEN

HUM-MOLGEN is a general interest communication list in human molecular genetics. Communication on HUM-MOLGEN is divided into TOPICS, from which subscribers can choose to receive mail or not. Topic-communication includes calls for collaboration, job opportunities, announcements, information about books or journals to appear in hard copy, discussion and monthly mailings about advances in topics of general human molecular genetics interest. Subscribers to the list include students, clinicians, molecular biologists and geneticists. Current editorial policy is to transfer all messages to the list, unless they are outside the scope of HUM-MOLGEN or clearly not of interest to all subscribers. A limited number of commercial messages are allowed on HUM-MOLGEN, but only when the content of these messages is of clear interest to all subscribers. Subscription is free. To subscribe, send the following email message: subscribe HUM-MOLGEN first_name last_name, to LISTSERV@nic.surfnet.nl and instructions will follow automatically. HUM-MOLGEN is now also available on the WWW (http://www.informatik.uni-rostock.de/HUM-MOLGEN/). The current editors of HUM-MOLGEN are Frank Zollmann and Arthur Bergen (owner). Further information is available by email (Bergen@amc.uva.nl or Zollmann.1@osu.edu).

BOOK REVIEWS

RNA cook books

RNA Processing Vols 1 and 2

by S.J. Higgins and B.D. Hames

Oxford University Press, 1994. £19.50 each pbk (232/278 pages)
ISBN 0 19 963343 6/0 19 963473 6

The past decade has proved to be highly memorable for those studying the RNA molecule. Its original role as a simple message from gene to protein (mRNA) or as an auxiliary molecule to aid in this process (tRNA and rRNA) has now evolved into a far more sophisticated picture. For example, RNA may display enzymatic activity as a ribozyme or possess a key regulatory function in gene expression, as exemplified by the alternative RNA-splicing patterns seen in the genes for sex determination in *Drosophila*. Most RNA molecules appear to go through often complex modification reactions, adjusting the original gene transcript to a final RNA structure with many different properties. Both the discovery of ribozymes and splicing have been marked by Nobel prizes, and the general level of scientific activity in RNA research has mushroomed from a few isolated research groups to many hundreds. The annual meeting on RNA processing has outgrown most possible conference venues and a new society has been spawned to lobby for and publish research into RNA.

With such a powerful wave of interest in RNA, it is highly opportune that IRL Press at Oxford University Press has just published two volumes on RNA-processing methodology in its well-known and appreciated *Practical Approach* series. Clearly, these books are not intended for bedtime reading but will prove to be invaluable 'cook books' for all experiments involving RNA. It is hard to fault the choice of topics covered or the authors prevailed upon to write them. One of the more often encountered difficulties in modern research is that the methods sections of papers become more and more skimpy. Even though all experiments published in respectable