Genetwork

BLAST, Blitz, BLOCKS and BEAUTY: sequence comparison on the Net

Sequence comparison is a powerful method for learning about the function, structure and evolution of newly identified genes and their encoded proteins. Every published DNA or protein sequence is compared to every other, if not by the authors, then automatically for inclusion in the Entrez database[1]. However, as with sequence retrieval (*Trends Genet.* 11, No. 6, 247–248), sequence database searching requires large computers, current databases in appropriate formats and specialized software. For these reasons, sequence searching was one of the first applications to make use of the Internet.

Straightforward searching

By far the most popular site for network sequence-database searching is the National Center for Biotechnology Information (NCBI), which currently has more than 8000 requests per day. NCBI provides a wide selection of databases, including non-redundant, comprehensive nucleotide and protein sequence databases, which are updated nightly. All searches at NCBI use the BLAST family of programs and information about all options is available from the NCBI BLAST Notebook[2]. Use of the NCBI server can be astonishingly fast, not only because BLAST is an inherently efficient algorithm, but also because it runs on machines especially configured to do the sequence searching. A query run at NCBI will not only search complete and current databases, but might even be quicker than the same search run on a local computer.

There are several ways of performing searches at NCBI, including via email – for information on this method, send a message to blast@ncbi.nlm.nih.gov containing the word HELP. Equally popular are network clients[3], programs that run on any computer and act as though they were doing the searches entirely locally, while actually doing all the work at NCBI. For users who have done sequence searches many times in the past, these provide a comfortable way to move towards use of the network. Indeed, now that the GCG Wisconsin Sequence Analysis package provides seamless integration of BLAST over the network, some biologists may be using the Net without even realizing it. For new users, probably the easiest method of sequence searching at NCBI is via the WWW[4]. Setting parameters and entering data this way is easy (though the sequence has to be cut and pasted), and hypertext results make browsing a breeze.

Standard search methods, other than BLAST, can be run at the European Bioinformatics Institute[5], Oak Ridge National Laboratory[6] and EERIE[7]. The most popular of these are FASTA, which predates BLAST but continues to be enhanced, and the Smith–Waterman algorithm. The Smith–Waterman method, also called Blitz or SSEARCH, might provide the best pairwise comparison results (Box 1), but is so computationally intensive that it is not practical to run on most computers. For this reason, EBI and ORNL run comparisons on massively parallel computers, which dramatically reduce search time.

Scientific showcases

Many researchers have developed special sequence comparison systems, and rather than distributing their programs and databases, instead make them available via email[8] on the WWW. These 'boutique' sequence comparison sites provide special resources available nowhere else and typically showcase a particular technology.

For example, it is possible to explore data from the *Caenorhabditis elegans* genome project[9] before this information has been fully analyzed and submitted to the major sequence databases. Also, at the same site, one can perform a similarity search against ProDom, an automatically generated database of protein

Box 1: Programs, Parameters and PAMs...Oh My!

The plethora of different systems for conducting sequence searches can become an embarrassment of riches; it is neither obvious where to start nor when to finish. While a comprehensive explanation of all the programs is well beyond the scope of this column, a good rule of thumb is to start with BLAST because it is very fast and may provide all the answers that are needed. FASTA is somewhat slower, and Smith–Waterman much more so. Bill Pearson has found that 'the new FASTA is significantly better than BLASTP' and for some sequences Smith–Waterman is better yet[18]. However, he also notes that, 'the slower programs rarely reveal significant relationships that the faster programs missed'. When none of these pairwise methods yields useful results, a pattern database search may prove fruitful. Even when using a particular method, the huge number of parameters can be daunting, and the particular choice of program might itself be confusing. For example, BLAST comes in some six flavours[19]; of these, BLASTP compares a protein sequence against a protein database and BLASTN compares a nucleotide sequence against a nucleotide database. It is typically best to start with the default parameters and matrices (e.g. PAMs and BLOSUMs) chosen by the program's authors, and then adjust them later if necessary. Much more information is provided in the very useful on-line discussions of searching by Geoff Barton and Keith Robison[20]. One final note: when reporting the results of any search, record precise details of both the methods and the databases used.

domains[10]. Another searchable domain database, Sbase, contains fewer sequences but is 'exhaustively annotated'[11]. A third set of databases of conserved domains is searched by a specialized version of BLAST, called BEAUTY. These can be accessed from the Baylor College of Medicine Search Launcher[12], a hypertext page created by Randall Smith, which reproduces on the WWW the palette of sequence searching sites that msu[13] provided for email. Also available from the Launcher are BLASTPAT and FASTPAT, which compare a query sequence with a database of homology-derived patterns (rather than actual sequences). Pattern-based searches are particularly useful when trying to identify distant members of a large family, and are also available for PROSITE[14] and BLOCKS[15].

A different dimension of information is provided by the similarity searches in the scop database[16]. A match gives links providing evolutionary and structural context for the query sequence and highlights the regions of similarity in a 3-D view of the homologous protein whose structure is known. Swiss-Model[17] also uses sequence searches to find proteins of known structure that are homologous to a query sequence. It then uses sophisticated algorithms to generate a model of the structure of the protein. These models should be used with great care: the fine details are certainly wrong, and even parts of the gross structure can be incorrect. Nonetheless, these methods provide an invaluable starting point for interpreting experimental knowledge in a structural context.

The tremendous quantities of information pouring out of genome laboratories are making sequence comparison ever more important and valuable, but the deluge of data means that many laboratories are finding it difficult to keep up. It is for this reason that email and WWW servers are crucial. Network-based sequence searching permits all biologists with network connections to share in the riches of sequence information, by placing the very latest and most exciting technology and data right on their desktop.

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

URLs

1 http://atlas.nlm.nih.gov:5700/Entrez/index.html
2 http://www.ncbi.nlm.nih.gov/Recipon/index.html
3 http://www.ncbi.nlm.nih.gov/Recipon/blast_network.html
4 http://www.ncbi.nlm.nih.gov/Recipon/blast_search.html
5 http://www.ebi.ac.uk/searches/searches.html
6 http://avalon.epm.ornl.gov/Grail-bin/EmptyGenquestForm
7 http://genome.eerie.fr/fasta/fasta-home.html
8 http://expasy.hcuge.ch/info/serv_ema.txt
9 http://www.sanger.ac.uk/~sjj/C.elegans_blast_server.html
10 http://www.sanger.ac.uk/prodom.html
11 http://www.icgeb.trieste.it/sbase/blast.html
12 http://kiwi.imgen.bcm.tmc.edu:8088/search-launcher/launcher.html
13 ftp://ftp.ebi.ac.uk/pub/software/unix/msu.tar.Z
14 http://expasy.hcuge.ch/sprot/scnpsite.html
15 http://www.blocks.fhcrc.org/
16 http://scop.mrc-lmb.cam.ac.uk/scop/
17 http://expasy.hcuge.ch/swissmod/SWISS-MODEL.html
18 http://www.techfak.unibielefeld.de/bcd/Curric/PrwAli/Lectures95/LectBPearson.html
19 http://www.ncbi.nlm.nih.gov/Recipon/blast_program.html
20 http://geoff.biop.ox.ac.uk/papers/rev93_1/rev93_1.html and http://twod.med.harvard.edu/seqanal/