Internet On-Ramp

BLAST...What Next?

Bruno Gaëta, eBioinformatics, Sydney, Australia

The vast amount of sequence information available through public and private sources is changing the face of biology. To access this information, researchers must become familiar with the computer tools available to analyze sequences, many of which are distributed on the Web. BLAST (1,2) allows investigators to identify sequences in a sequence database that are similar to a supplied sequence. However, this search is often only the first step in a more complex series of analyses in which the matches identified by the search are processed to yield additional information on the query sequence and its biological relationships.

Result Visualization

The first BLAST output processing programs added no information to the BLAST results but made them easier to view. The standard output consists of an often long list of matching database sequences, followed by pairwise alignments between the query sequence and each of the best matching database sequences. Browsing this long text file is made easier by the hypertext links between the database hits in the list and their corresponding pairwise alignments. The description of each hit can also be linked with the original database entry; both features are now standard for BLAST.

Some implementations of BLAST provide alternative views of the results, besides the standard output view. For example, the BLAST server at the National Center for Biotechnology Information (NCBI) (<http://www.ncbi.nlm.nih.gov>) offers the option to replace the series of pairwise alignments with one large multiple sequence alignment of all matching database sequences and the query sequence. The BLAST server provided by the National Center for Genome Resources (<http://seqsim.ncgr.org/newBlast.html>) saves the BLAST output as a tab- or comma-delimited text file that can be imported directly into a spreadsheet for display, filtering and further processing.

MSPcrunch (6) is another output processor designed for high-throughput analysis of large genomic sequences. Unlike programs that simply reformat the BLAST output, MSPcrunch filters the results and recalculates some of the scores to emphasize hits that may be biologically significant. It also filters out alignments of poor biological significance caused by regions of biased composition and combines small but significant ungapped alignments into a larger, more biologically relevant gapped alignment. The program also provides a "big picture" output option to represent the alignments graphically, together with a range of output formats suitable for entering into semiautomatic genome annotation tools and databases. MSPcrunch is available on the Web at the Pasteur Institute (<http://bioweb.pasteur.fr/seqanal/blast/intro-uk.html>) and the Telethon Institute of Genetics and Medicine (<http://www.tigem.it/LOCAL/BLAST/blast.html>). The program is also available at <http://www.embnet.se/new/mspup.html>, where users can enter their own BLAST outputs generated with other BLAST servers. Note that the current version of MSPcrunch is not compatible with some versions of BLAST, including the gapped BLAST version 2 available at NCBI, which already provides gapped alignments and a graphical display.

Multiple Sequence Alignments

The NCBI gapped BLAST interface makes it possible to reformat the standard output to produce a multiple sequence alignment of all matched sequences—either as full sequences or showing only the differences between the database sequences and the query sequence. However, this alignment is not in a format that can be input easily into other programs for further analysis and is therefore suitable mainly for display and visualization. On the other hand, MView (3), a multiple sequence alignment conversion, viewing and reformatting tool, can take a BLAST output and convert it into a multiple sequence in a range of formats for display on a Web page or for input into other analysis programs. MView is available on the Web through the Pasteur Institute Web server (http://bioweb.pasteur.fr/seqanal/interfaces/mview_blast.html) and has also been integrated into the PIR Integrated Sequence Analysis Environment (<http://www-nbrf.georgetown.edu/pirwww/search/pirisa.html>) for reformatting the output of BLAST searches of the PIR database. Integrated Web-based bioinformatics analysis systems such as BioNavigator (<http://www.BioNavigator.com>) allow the researcher to save the database hits so they can be entered directly into separate multiple sequence alignment programs such as ClustalW (7). Instead of using the alignments in the BLAST output, the alignment program recalculates the alignments and aligns the database sequences rather than just the matching sequence segments found by BLAST.

Phylogeny

Multiple sequence alignments are commonly used as data sets for the construction of phylogenetic trees. Using an integrated platform such as BioNavigator or the Pasteur Institute's PISE interface (<http://bioweb.pasteur.fr/intro-uk.html>) makes it easy to run a BLAST search and extract the matching database sequences, then align them to produce phylogenetic trees. (A number of programs and methods are available on different Web sites for constructing the phylogenetic trees from the alignment.) Note that all tree-building programs make certain assumptions and that using these programs without being aware of their underlying model can result in misleading phylogenetic trees. A BLAST server has also been set up at <http://www.Bork.EMBL-Heidelberg.DE/Blast2e/> for differentiating orthologues (homologous genes or proteins in different species) and paralogues (homologues in the same species that are the result of gene duplication) using a phylogenetic approach. This server performs a BLAST search and offers the option of creating a multiple sequence alignment from the results and deriving a phylogenetic tree from the alignment. The tree can be compared to the known phylogeny of the species from which the sequences were obtained. In principle, a phylogenetic tree constructed from orthologous sequences should be identical to the species tree.

Profiles and Hidden Markov Models

A multiple sequence alignment can also be used to construct a mathematical model such as a sequence profile (also called position-specific scoring matrix or PSSM) or a hidden Markov model (HMM) of the group of sequences in the alignment. If the sequences all belong to the same protein family, the model can be used as a representation of this family and can be used to search a sequence database for new

Internet On-Ramp

members of the family or to test if a new sequence belongs to the family. Searching for new members of a sequence family with a profile or an HMM is more sensitive than searching with an individual member of this family because the model brings together information from many sequences and is less affected by individual variation in sequences.

The easiest way to use profiles for database searching is to make use of the program PSI-BLAST (position-specific iterated BLAST) (2) developed at NCBI. PSI-BLAST starts with a standard blastp search of a protein database, and sequences with sufficient similarity to the query are then aligned together and used to build a profile to search the database again for new matching sequences. The new matching sequences can be added to the profile and the search repeated until no new sequences are identified. PSI-BLAST is available from an increasing number of Web sites including NCBI, BioNavigator and the DNA database of Japan (<http://crick.genes.nig.ac.jp/homology/welcome-e.shtml>). PSI-BLAST at NCBI also allows the program to be run interactively, with users selecting sequences to be added to the profile after each round of searching. It is also possible to perform the same type of analysis manually, using a different type of model for the sequence family. An integrated platform makes it much easier to go from a BLAST output to a multiple sequence alignment. It then creates the model from the multiple sequence alignment and finally searches a database using the model to identify new members of the protein family. For example, BioNavigator provides a Web front end to the HMMer package (4), which builds HMMs from multiple sequence alignments and then uses them to search sequence databases.

Sequence Motifs

Many sequence families, domains and functions can be identified using specific combinations of conserved residues that form recognizable patterns or motifs. Motif databases, such as PROSITE (available at <http://www.expasy.ch>), collect protein motifs to facilitate sequence identification. Searches for sequence patterns can yield information complementary to that provided by a BLAST search, and the two approaches can be combined when identifying the function of new proteins.

BEAUTY (8) is a program based on BLAST that allows the researcher to identify protein sequences using both sequence similarity information and information about conserved patterns and motifs present in the query sequence and the matching database sequences. The program performs a BLAST search of databases of conserved protein regions and motifs and reports which regions of the query sequence matched known motifs and conserved domains. Investigators can then decide whether or not the query sequence contains all the conserved sequence patterns required for performing a particular biological function and can provide a much stronger indication of the function of a protein or a domain than simple sequence similarity. BEAUTY is available through the Baylor College of Medicine Search Launcher (<http://dot.imgen.bcm.tmc.edu:9331/seq-search/protein-search.html>).

Another BLAST variant combining sequence similarity and searches for specific motifs is PHI-BLAST (pattern hit initiated BLAST) (9) available through the NCBI Web site. PHI-BLAST requires as input both a query sequence and a particular sequence pattern present in this sequence. PHI-BLAST searches for sequences containing the input sequence pattern that

are similar to the query sequence around this pattern. This approach is more specific than just searching for the conserved motif, which may match many false positives in the database.

Protein Structure

A common approach to predict the structure of a protein from its sequence is to use the sequence as query in a search of a database of proteins of known structure. The structure of matching sequences can then be used as a guide for creating a model of the structure of the unknown protein. This approach works well when the query protein has close homologues of known structure; it is the basis for the SWISS-MODEL (5) 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 prediction. BioNavigator 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, then enter the query sequence and its matching structure into other programs for building a structural model of the query protein.

BLAST is most commonly used as a stand-alone program but can also be the basis for more involved sequence analysis. This analysis can be carried out using dedicated programs such as BEAUTY and MSPcrunch or by extracting the database hits identified by the BLAST search and entering them into other sequence analysis programs. Integrated bioinformatics packages such as BioNavigator greatly facilitate the latter task compared with standard BLAST interfaces available on the Web, which require users to copy and paste or download individual hits one by one. These integrated services also offer a large collection of different methods for analyzing the sequences identified by BLAST.

REFERENCES

1. Altschul, S.F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.
2. Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.
3. Brown, N.P., C. Leroy and C. Sander. 1998. MView: a Web compatible database search or multiple alignment viewer. *Bioinformatics* 14:380-381.
4. Eddy, S. 1998. HMMER User's Guide. Biological sequence analysis using profile hidden Markov models. <http://hmmerr.wustl.edu/>
5. Guex, N., A. Diemand and M.C. Peitsch. 1999. Protein modelling for all. *Trends Biochem. Sci.* 24:364-367.
6. Sonnhammer, E.L. and R. Durbin. 1994. A workbench for large-scale sequence homology analysis. *Comput. Appl. Biosci.* 10:301-307.
7. Thompson, J.D., D.G. Higgins and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.
8. Worley, K.C., B.A. Wiese and R.F. Smith. 1995. BEAUTY: an enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results *Genome Res.* 5:173-184.
9. Zhang, Z., A.A. Schäffer, W. Miller, T.L. Madden, D.J. Lipman, E.V. Koonin and S.F. Altschul. 1998. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* 26:3986-3990.