Visual BLAST and Visual FASTA: graphic workbenches for interactive analysis of full BLAST and FASTA outputs under Microsoft Windows 95/NT

P. Durand, L. Canard and J. P. Mornon

Abstract

Motivation: When routinely analysing protein sequences, detailed analysis of database search results made with BLAST and FASTA becomes exceedingly time consuming and tedious work, as the resultant file may contain a list of hundreds of potential homologies. The interpretation of these results is usually carried out with a text editor which is not a convenient tool for this analysis. In addition, the format of data within BLAST and FASTA output files makes them difficult to read.

Results: To facilitate and accelerate this analysis, we present, for the first time, two easy-to-use programs designed for interactive analysis of full BLAST and FASTA output files containing protein sequence alignments. The programs, Visual BLAST and Visual FASTA, run under Microsoft Windows 95 or NT systems. They are based on the same intuitive graphical user interface (GUI) with extensive viewing, searching, editing, printing and multithreading capabilities. These programs improve the browsing of BLAST/FASTA results by offering a more convenient presentation of these results. They also implement on a computer several analytical tools which automate a manual methodology used for detailed analysis of BLAST and FASTA outputs. These tools include a pairwise sequence alignment viewer, a Hydrophobic Cluster Analysis plot alignment viewer and a tool displaying a graphical map of all database sequences aligned with the query sequence. In addition, Visual Blast includes tools for multiple sequence alignment analysis (with an amino acid patterns search engine), and Visual FASTA provides a GUI to the FASTA program.

Availability: The programs are freely available on the Web (http://www.lmcp.jussieu.fr/~durand).

Contact: E-mail: durand@lmcp.jussieu.fr

Introduction

Computer-assisted analysis of protein sequences is extensively used in sequence database searches. The central goal of such searches is the rapid identification of the function(s) and/or tertiary structure of a new protein by analogy with a protein of known function and/or three-dimensional (3D) structure. Currently, high-speed sequence comparison methods, such as FASTA (Pearson and Lipman, 1988) and BLAST (Altschul et al., 1990), are the most commonly used sequence database-searching programs. Depending on the scoring matrices and gap penalties used, these two sequence comparison algorithms produce complementary information. Indeed, although BLAST and FASTA globally use the same strategy to scan the databases, there are two major differences between them. First, FASTA provides optimized gapped sequence alignments, whereas BLAST provides ungapped sequence alignments. Second, the two programs do not use the same scoring system to rank the database sequences matched with the query sequence. It is, therefore, of interest to exploit the results of both programs in a database search strategy (Pearson, 1995). BLAST and FASTA report results of the searches in text files which contain a list of brief descriptions of statistically significant matching database sequences followed by the alignments of the query sequence with each of the matched sequences. With the present large-scale genome sequencing projects, the amount of publicly available DNA and protein sequences is rapidly increasing. Thus, detailed analysis of a database search can be extremely time consuming and could rapidly become quite tedious, especially for inexperienced users, as the resultant file may contain a list of hundreds of potential homologies. Currently, the analysis of BLAST and FASTA results requires a text editor or, less frequently, a printed copy of the BLAST/FASTA output files. However, a text editor is not a convenient tool for this analysis because it implies a sequential reading of the BLAST/FASTA files and a manual localization of sequence alignments within these files.

We have developed two programs, Visual BLAST and Visual FASTA, specially designed to analyse full BLAST and FASTA outputs. These tools have been created to propose an alternative to the text editor, and to simplify and improve the interpretation of BLAST and FASTA results. Both programs benefit from the Windows 95/NT graphical
user interface (GUI) to display the results in a graphical and more convenient manner than in the original BLAST and FASTA output files. The GUI offers the user a new, interactive and rapid method by which he views and manages all the data reported in the results. These programs implement and automate in silico a manual methodology used for detailed analysis of database search results made with FASTA and BLAST. This methodology combines the analysis of pairwise sequence alignments, the 2D plot comparisons based on Hydrophobic Cluster Analysis (HCA) (Gaboriaud et al., 1987), the analysis of multiple sequence alignments and the search for amino acid patterns. These tools are especially useful for the identification of protein families and the localization of conserved patterns between related protein sequences. In addition, Visual FASTA exploits the power of the Windows 95/NT multi-threading operating system: the user can execute a FASTA database search directly from Visual FASTA. The search is executed as a background process and the results are automatically displayed within Visual FASTA.

System and methods

Visual BLAST and Visual FASTA implement standard Windows features such as on-line help, dialog boxes, toolbars and high-quality printing capabilities. Their use is optimized by extensive use of the mouse to access all commands of the software via either pulldown menus or toolbars. Visual BLAST and Visual FASTA are available for i486 and Pentium computers running Microsoft Windows 95 or NT. The minimum system configuration is the one necessary for Windows, but with two remarks. First, the minimum screen resolution is 800 × 600, but 1024 × 768 is highly recommended to take advantage of the graphic capabilities of the programs. Second, it is recommended that the computer be connected to the Internet to retrieve BLAST/FASTA output files produced by the various BLAST/FASTA database search services available via the Internet. Finally, to execute a FASTA database search from Visual FASTA, one can install on the local computer the original FASTA program written by W. Pearson (Pearson and Lipman, 1988) as well as the database(s). (The FASTA program is available on the anonymous ftp server of the University of Virginia at ftp.virginia.edu in the/pub/fasta/dos directory.)

Implementation of Visual BLAST and Visual FASTA

Data acquisition and presentation

Visual BLAST is able to read data from the various releases of blastp, blastx and tblastn programs available in the GCG package (Devereux et al., 1984), the BLAST Web servers at the NCBI (http://ncbi.nlm.nih.gov/) and ExPASy (http://expasy.hcuge.ch/) and the NCBI's Network BLAST client software (Network BLAST is available by anonymous ftp at ncbi.nlm.nih.gov in the /blast/network/blast2/win32 directory). Visual FASTA is able to read data from the FASTA provided with the GCG package and from the original FASTA of W. Pearson (see System and methods). Both programs read only the program introduction and the alignment sections of the BLAST/FASTA output file to retrieve database information as well as sequence alignments and statistical data. All this information is displayed within the main window of the programs. The main window of Visual BLAST (or Visual FASTA) improves the presentation of the results via a new organization of the data. In a BLAST (or FASTA) output, the list of one-line summaries gives only partial information about a specific alignment. To obtain the complete information (statistics, full database sequence name and sequence alignment), the user has to locate it manually within the alignment section of the BLAST (FASTA) output. With Visual BLAST (or Visual FASTA), the list of one-line summaries and the alignment section information are displayed simultaneously. To do this, their main window is divided into four windows. The first window is a listbox, called the Browser, where each entry is numbered and displays the database sequence identifier as well as the scores used by BLAST (or FASTA) to rank the results. The other three windows display information corresponding to the currently selected Browser entry: the database sequence full name and length, the statistical information and the sequence alignment. For Visual BLAST only, the number of HSPs (high-scoring segment pairs) for each database sequence is also clearly shown at the bottom of the Browser. In contrast to a Text Editor, the Browser of Visual BLAST and Visual FASTA allows a rapid and interactive survey of the BLAST results: a single click with the mouse on a Browser entry displays all the information about this alignment.

Analytical tools: ID viewer, String Finder, HCA viewer, BLAST(FASTA)-map, MulSeq Editor and Find patterns

Within BLAST and FASTA output, sequence alignments are in black and white and are truncated in order to fit in the available space of the file. Our programs use the 1D viewer to display full linear and colour pairwise sequence alignment. This viewer is implemented within the main window of both Visual BLAST and Visual FASTA. Each sequence alignment can be printed directly on any printer device supported by Windows. All the other tools are directly available from the main window of Visual BLAST and Visual FASTA.

String Finder allows the user to search the entire BLAST/FASTA output file for any character strings. String Finder reports all the occurrences of the keyword at a time, whereas the 'Find Text' function of a Text Editor, frequently used to analyse BLAST or FASTA files, can report only one occurrence at a time.
HCA viewer is a colour 2D pairwise alignment viewer based on the HCA representation (Gaboriaud et al., 1987). We currently use the algorithm described by Henrissat et al. (1990) to create the graphical representation of the HCA plots. The implementation of HCA viewer automates the classical and manual HCA plot comparison which requires the following steps: (i) recovery of the sequences of the 1D alignment in the databases; (ii) use of a specific program to produce, sequence by sequence, the 2D HCA plots; and (iii) manual reconstitution of the HCA plot alignment. In this way, and for the first time, HCA has been directly associated with database search results analysis and allows researchers to use the HCA methodology more easily (for a review of HCA methodology, see Lemesle-Varloot et al., 1990). The Browser then allows the user to instantaneously see 1D and 2D sequence alignments in the corresponding viewer. In addition, HCA plot alignment can be printed directly on any printer device supported by Windows.

BLAST (or FASTA)-map displays a graphical map which shows the position within the query sequence of each database sequence matched with it. Such a map is unavailable with the native BLAST (or FASTA) outputs. The goal of BLAST (or FASTA)-map is to automate the manual localization of one (or more) region(s) of the query sequence where many database sequences overlap. Thus, this map can be very useful to locate conserved regions between related sequences.

Only Visual BLAST implements tools to analyse multiple sequence alignments (MSA) piled up from the ungapped alignments reported in BLAST outputs: a complete MSA editor (MulSeq Editor) and an amino acid pattern search engine (Find patterns). This allows the user to analyse a MSA to locate conserved regions of related protein sequences and/or to prepare a database profile scanning (Gribskov et al., 1987). The MSA can be saved in the MSF format readable by several programs from the widely used GCG package (Devereux et al., 1984).

Fig. 1. The main window of Visual BLAST with HCA viewer and String Finder. (All figures shown here are black and white printed plates of a 1024 x 768 colour screen copy.) Example of Visual BLAST working session showing the main window (bottom), HCA viewer (top left) and String Finder (top right). Within the Browser (the listbox located on the left of the main window), we have selected the 24th entry. All the information about this entry (sequence name and alignment, as well as statistical data) is displayed within all the other windows of the Visual BLAST main window. Sequence alignments are displayed using the HCA colour codes (Lemesle-Varloot et al., 1990) with the query sequence above the database sequence.
Database searching tool: VF-Job

Because the FASTA program as well as the databases are available for personal computers, we have implemented a GUI to the FASTA program: VF-Job. It consists of a dialog box where the user specifies the parameter settings of the FASTA search. In this way, the user no longer works with complex command-line. VF-Job can be invoked alone or from Visual FASTA. The FASTA search is executed as a background process, and when it is completed VF-Job displays the results in a new Visual FASTA window.

Example

To illustrate the use of Visual BLAST (the use of Visual FASTA is similar to that of Visual BLAST), we summarize the principles of analysis of the BLAST search of an enzyme belonging to the glycosyl hydrolase group. We have searched...
the NCBI’s non-redundant protein database using BLASTP and the bovine lysosomal β-mannosidase (PIR:A55881; Chen et al., 1995) as the query sequence.

The BLAST output is displayed within the main window of Visual BLAST (Figure 1). Before a detailed analysis of the sequence alignments, we have displayed a graphical map of the BLAST file using BLAST-map. As shown in Figure 2a, this map reveals that many dozen database sequences match with a same region of the query sequence. This automatic survey of the BLAST file rapidly orients further analysis towards this region of the query sequence. Then, using the Browser, we have identified several β-galactosidases and β-glucuronidases in the first 20 entries of the list. To identify all other such enzymes, we used String Finder to search the entire BLAST file for all occurrences of the keyword beta (Figure 1). String Finder has found about three dozen β-galactosidases and β-glucuronidases matched with the query sequence. However, the probability score (from 0.011 to 1.0) makes these similarities questionable. Nevertheless, careful survey of BLAST output using HCA viewer shows conservation of the hydrophobic clusters and similar residues between a same region of the query sequence (from amino acid 420 to amino acid 460) and the β-galactosidases and β-glucuronidases reported in the BLAST file (Figure 1). A detailed analysis of all these alignments has been carried out using MulSeq Editor (Figure 2b). In particular, the consensus sequence reveals the presence of several highly conserved amino acids. Using Find patterns, we have then searched all the sequences displayed within the MulSeq Editor for glycosyl hydrolase patterns referenced in the Prosite database (Bairoch, 1992). This search reveals that many of these sequences contain a family 2 glycosyl hydrolase pattern (Figure 2c) which is located within the conserved region highlighted by BLAST-map and the MulSeq Editor consensus sequence. This pattern includes the hydrophobic amino acids of a β strand conserved in the entire protein family and two strictly conserved residues: an asparagine followed by a glutamate (the two last residues of the consensus sequence in Figure 2b) involved in the catalytic mechanism of the family 2 glycosyl hydrolases. This allows...
us to conclude that the bovine $\beta$-mannosidase is a new member of the glycosyl hydrolase family 2. However, this homology has not been previously reported due to three single conservative substitutions preceding the glutamate catalytic residue in the pattern. Thus, a single search for this pattern within $\beta$-mannosidase makes it unsuccessful and highlights the advantage of searching for patterns within a multiple sequence alignment of related sequences. This can be done with Find patterns. A more complete analysis of the $\beta$-mannosidase has been performed to support our conclusion (Durand et al., 1997).

Discussion

The study of a new protein often begins with a database similarity search. For many years, FASTA and BLAST have been the most widely used database-searching programs. The results files provided by these programs quite often contain hundreds of sequence alignments, and researchers are confronted with the interpretation of an enormous amount of data. Detailed analysis of BLAST and FASTA outputs is nonetheless important to identify a group of related sequences or to locate conserved sequence motifs. As the results files of BLAST and FASTA are not easy to interpret, especially for inexperienced users, several authors have developed additional programs which use specific algorithms to treat FASTA and BLAST outputs. Examples are FTHOM (Hegyi and Pongor, 1993), BLA (Tatusov and Koonin, 1994), BEAUTY (Worley et al., 1995) and MulBlast (Labesse, 1996). The first three programs use information contained within SwissProt or Prosite databases to identify domains or patterns with or without the generation of a multiple sequence alignment. This concept limits these programs to the identification of already known protein domains or patterns. This limitation is overcome by MulBlast, recently developed in our laboratory, which succeeds in detecting any conserved segments between selected proteins reported in a BLAST output file. However, neither these programs nor, to our knowledge, any previous ones, have targeted a complete and friendly interactive analysis of the full BLAST and FASTA information by a combination of the use of 1D sequence alignment analysis, HCA and multiple sequence alignment analysis. This was our goal with Visual BLAST and Visual FASTA.
Indeed, we have created interactive GUI-oriented tools designed for detailed analysis of information contained in sequence database search outputs. In particular, we have centred our project to enhance the presentation of database-searching results and to combine several tools which allow the user to manipulate the data as needed. To do so, Visual BLAST and Visual FASTA have been developed to be interactive, easy-to-use and productive tools for inexperienced and infrequent users as well as experienced researchers in the interpretation of BLAST/FASTA search outputs. Experienced users have additional tools designed for HCA and, in Visual BLAST, for the analysis of multiple sequence alignments in order to locate sequence motifs shared by a group of related proteins or to prepare data for database profile scanning.

For future versions of our analysis package, we plan to merge Visual BLAST and Visual FASTA in order to combine and manage BLAST and FASTA results within a single program. The Visual BLAST interface used in conjunction with NCBI’s Network BLAST client software, as well as Visual FASTA used with VF-Job, provide powerful environments for the execution and interpretation of BLAST/FASTA database searches.

The executables of Visual BLAST and Visual FASTA, as well as a complete HTML documentation, are available on the Web at the URL http://www.lmcp.jussieu.fr/~durand/ in the section Softwares.

Note: a general HCA plot computing facility is also available in our laboratory at the URL http://www.lmcp.jussieu.fr/~mornon. This service is called DRAWHCA.

Acknowledgements

We thank Isabelle Callebaut and Nathalie Colloc’h for helpful discussions and critical reading of the manuscript. This work was supported by the Association Vaincre les Maladies Lysosomales (VML) (Evry, France) through a PhD grant to P.D.

References


Received on November 25, 1996; revised on January 29, 1997; accepted on January 30, 1997