BIOINFORMATICS APPLICATIONS NOTE

Vol. 15 no. 5 1999
Pages 422–423

WebBLAST 2.0: an integrated solution for organizing and analyzing sequence data

Erik S. Ferlanti, Joseph F. Ryan, Izabela Makalowska and Andreas D. Baxevanis

Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Building 49, Room 2C-76, Bethesda, MD 20892-4431, USA

Received on November 5, 1998; accepted on February 3, 1999

Abstract

Summary: *WebBLAST is a suite of programs intended to assist in organizing sequencing data and to provide first-pass sequence analysis in an automated fashion. Data processing is fully automated, with end-users being presented both graphical and tabular summaries of data that can be viewed using any Web browser.*

Availability: *The program is free and available at http://genome.nhgri.nih.gov/webblast.*

Contact: *webblaster@nhgri.nih.gov*

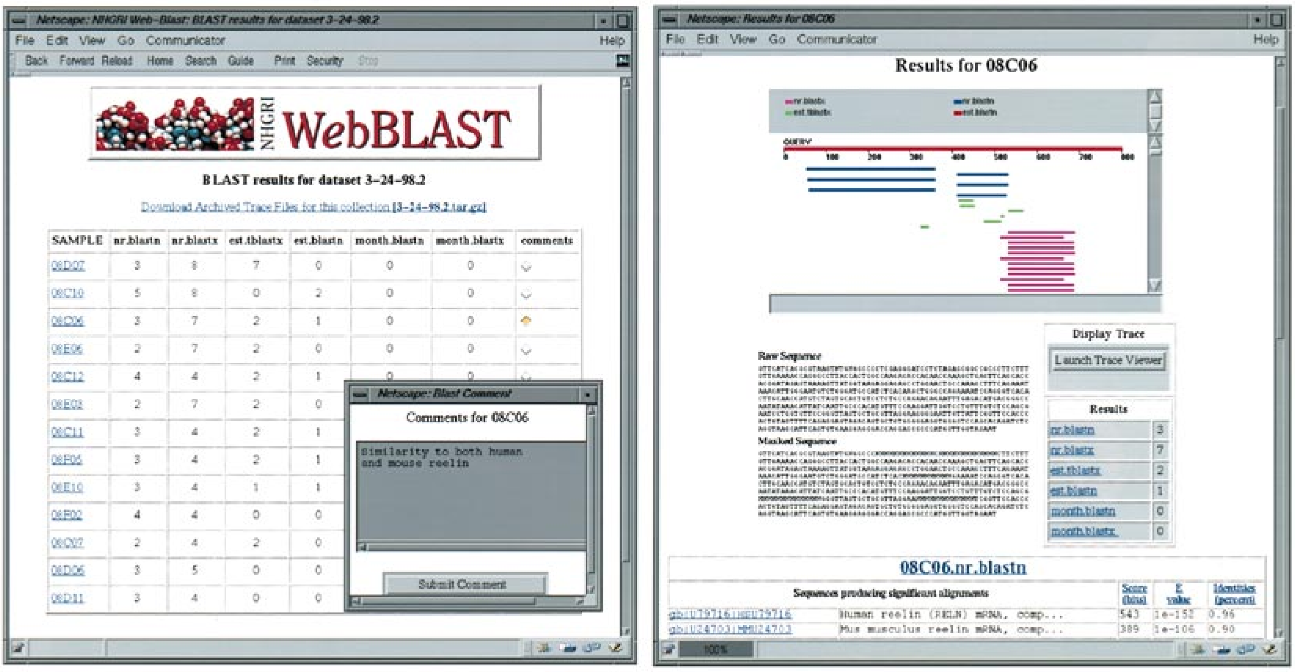
As the numbers of positional cloning projects, systematic organismal sequencing projects and other sequencing efforts related to the Human Genome Project advance at breakneck speed, there is an ever-increasing need for automated methods that will both archive sequencing data and perform basic analysis tasks. In response to the lack of freely available programs to fulfill this need, we have developed WebBLAST, a suite of Macintosh and UNIX-based programs that can collect and organize sequence data, and provide first-pass sequence analysis in the form of BLAST searches (Altschul *et al.*, 1990, 1997). WebBLAST will provide investigators with a tool through which they can develop an effective strategy for both data handling and elementary analysis in the early stages of the maturation of sequence data.

One of the main features of WebBLAST is that its system requirements are minimal. The program requires only MacPerl on the Macintosh client, Perl 5 and BLAST 2.0 on the UNIX client, and a Web server. Since we anticipate the use of this program in smaller positional cloning laboratories who do not necessarily have high-level computational support, freeware products were purposefully chosen over commercial relational database products. By doing so, both costs and administrative overhead are kept to an absolute minimum.

Data enter the WebBLAST pipeline through the use of a MacPerl droplet. After a sequencing run is completed, the sequencing software creates a folder containing both sequence and trace files from that run. This newly created folder is simply dragged onto the droplet application (SequenceUpload) to initiate the transfer. Files are then transferred to a holding directory on the UNIX server. All data are stored on the UNIX server in a filesystem database under the document root of the Web server. Typically, each sequencing effort is given its own project, and each BAC or YAC clone within that sequencing effort is designated as a subproject. Each sequence directory contains the sequence, the trace for that sequence, and BLAST reports, amongst other data.

A nightly cron job invokes the processing of new data previously uploaded using the MacPerl droplet. Sequences in the holding directory are first moved into the appropriate position within the filesystem hierarchy. Sequences are then screened for the presence of vector, alu and other repetitive sequences, and masked. Local BLAST searches are then run on the masked data, and the results are written to the database. Configuration files allow users to specify which BLAST programs are called, which databases are used for the BLAST queries, and any command-line options that are to be passed to BLAST. The resulting BLAST reports are parsed, and high-scoring segment pairs that meet the statistical cut-off are written to disk as flatfiles. Finally, the routines may be set up so that collections of data are re-BLASTed once a month against the NCBI month database in the same fashion as described above, thereby ensuring that BLAST results are kept up to date.

Access to the data is through a Web front-end, allowing users access to the data regardless of platform. Upon selecting a collection, a summary table of significant hits for all sequences within that collection is displayed (Figure 1A). From the summary table, users can select an individual sequence, which produces a detail page for that sample (Figure 1B). A Java applet at the top of the detail page gives a graphical view of the distribution of BLAST hits, with each BLAST hit being represented by a color-coded bar. The raw sequence and masked sequence are shown below the graphical view, along with a summary table for each type of BLAST run. These are, in turn, followed by sorted tables indicating the individual hits produced by each BLAST run. Clicking onto an accession number within these tables will also move the user directly to the relevant alignment in the BLAST report, and accession numbers within the BLAST report are hyperlinked to NCBI Entrez (Baxevanis, 1998).


Netscape: NHGRI Web-Blast: BLAST results for dataset 3-24-98.2
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
NHGRI WebBLAST

BLAST results for dataset 3–24–98.2

Download Archived Trace Files for this collection [3–24–98.2.tar.gz]

SAMPLE	nr.blastn	nr.blastx	est.tblastx	est.blastn	month.blastn	month.blastx	comments
08D07	3	8	7	0	0	0	○
08C10	5	8	0	2	0	0	○
08C06	3	7	2	1	0	0	●
08E06	2	7	2	0	0	0	○
08C12	4	4	2	1	0	0	○
08E03	2	7	2	0			
08C11	3	4	2	1			
08F05	3	4	2	1			
08E10	3	4	1	1			
08F02	4	4	0	0			
08C07	2	4	2	0			
08D06	3	5	0	0			
08D11	3	4	0	0			

Comments for 08C06

```
Similarity to both human
and mouse reelin
```

Submit Comment
Netscape: Results for 08C06
File Edit View Go Communicator Help

Results for 08C06

Results	
nr.blastn	3
nr.blastx	7
est.tblastx	2
est.blastn	1
month.blastn	0
month.blastx	0

08C06.nr.blastn

Sequences producing significant alignments		Score (bits)	E value	Identities (percent)
gb|U79716|HSU79716	Human reelin (RELN) mRNA, comp...	543	1e-152	0.96
gb|U24703|MMU24703	Mus musculus reelin mRNA, comp...	389	1e-106	0.90

Fig. 1. WebBLAST views generated by the web-blast.cgi program. (**A**) Summary page for an individual data set. A comment pop-up window is shown in the inset. Full comment information is visible in either the pop-up window or on the detail page. (**B**) Detail page for an individual sequence. Note the Java graphical viewer, which allows users to inspect HSPs visually and directly jump to the relevant parts o f the text output. The graphical viewer shows all HSPs, while the actual hit counts shown in the tables represent the absolute number of sequences found. Table columns may be sorted by clicking on any column header. A WebBLAST demonstration site using sample data is available on-line at http://genome.nhgri.nih.gov/webblast/demo/web-blast.cgi.

As WebBLAST has been used in multi-user environments to date, each sample is given its own comment box for user annotation (Figure 1A). A simple search engine has been built into WebBLAST that allows users to search for individual sequences. In addition, users can also directly view sequence traces through the use of the Whitehead TraceViewer, a Java applet that can be invoked automatically (Zody, 1998). WebBLAST also provides for export to assembly programs such as the PHRED/PHRAP/CONSED suite (Ewing *et al*., 1998; Gordon *et al*., 1998) and the Staden package (Staden, 1996).

WebBLAST has been used in a number of mapping and positional cloning projects to date (e.g. The International FMF Consortium, 1997; Southard-Smith *et al*., 1998) and has proven helpful in both narrowing down critical regions and identifying promising candidates that may be implicated in the causation of human genetic or genomic disorders.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Baxevanis,A.D. (1998) Information retrieval from biological databases. In Baxevanis,A.D. and Ouellette,B.F.F. (eds), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley and Sons, New York, pp. 98–120.

Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*., **8**, 175–185.

Gordon,D., Abajian,C. and Green,P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res*., **8**, 195–202.

The International FMF Consortium (1997) Ancient missense mutations in a new member of the RoRet gene family are likely to cause familial Mediterranean fever. *Cell*, **90**, 797–807.

Southard-Smith,E.M., Collins,J.E., Ellison,J.S., Smith,K.J., Baxevanis,A.D., Touchman,J.W., Green,E.D., Dunham,I. and Pavan,W.J. (1999) Comparative analyses of the *Dominant megacolon-SOX10* genomic interval in mouse and man. Mammalian Genone, in press.

Staden,R. (1996) The Staden sequence analysis package. *Mol. Biotechnol*., **5**, 233–241.

Zody,M. (1998) Whitehead Java Trace Viewer. Whitehead Institute/MIT Center for Genomic Research, Cambridge, MA. ftp://xenon.wi.mit.edu/pub/traceviewer.distribution.tar.gz.