BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 8 2001
Pages 744–745

Multi-query sequence BLAST output examination with MuSeqBox

Liqun Xing[1] and Volker Brendel[1, 2]

[1]Department of Zoology and Genetics and [2]Department of Statistics, Iowa State University, Ames, IA 50011, USA

Received on October 25, 2000; revised on March 19, 2001; accepted on April 6, 2001

ABSTRACT

Summary: MuSeqBox is a program to parse BLAST output and store attributes of BLAST hits in tabular form. The user can apply a number of selection criteria to filter out hits with particular attributes. MuSeqBox provides a powerful annotation tool for large sets of query sequences that are simultaneously compared against a database with any of the standard stand-alone or network–client BLAST programs. We discuss such application to the problem of annotation and analysis of EST collections.

Availability: The program was written in standard C++ and is freely available to noncommercial users by request from the authors. The program is also available over the web at http://bioinformatics.iastate.edu/bioinformatics2go/mb/MuSeqBox.html.

Contact: vbrendel@iastate.edu

The ready availability of complete genomes necessitates computational approaches to functional genome annotation (Brenner, 1999). A starting point for any such analysis is gene identification and annotation on the basis of inferred homology to previously established genes and gene products. The NCBI BLAST database search tool (Altschul *et al.*, 1997) is probably the most popular program designed to solve single query problems. For genome-wide comparisons, the comparison challenge changes from the standard one-to-many to a many-to-many query problem, for example when comparing all proteins of one bacterial species to all proteins from another species (Koonin *et al.*, 2000). In this case, a first step analysis might involve successive applications of BLAST with each protein of one species in turn serving as a query against the database of all protein sequences of the second species.

Inspection and interpretation of what may be thousands of individual BLAST output files clearly must involve automated post-processing of the output (e.g. Sonnhammer and Durbin, 1994). We describe our MuSeqBox program that was developed for such applications. MuSeqBox serves as a filter that stores only essential features of BLAST hits in tabular form. Flexible ‘slicing and dicing’ capabilities allow post-processing of the hit table to select only hits with particular attributes. We first describe the general capabilities of the program and then discuss an exemplary application of annotation and interpretation of species-specific sets of Expressed Sequence Tags (ESTs).

TABULATED BLAST HIT INFORMATION

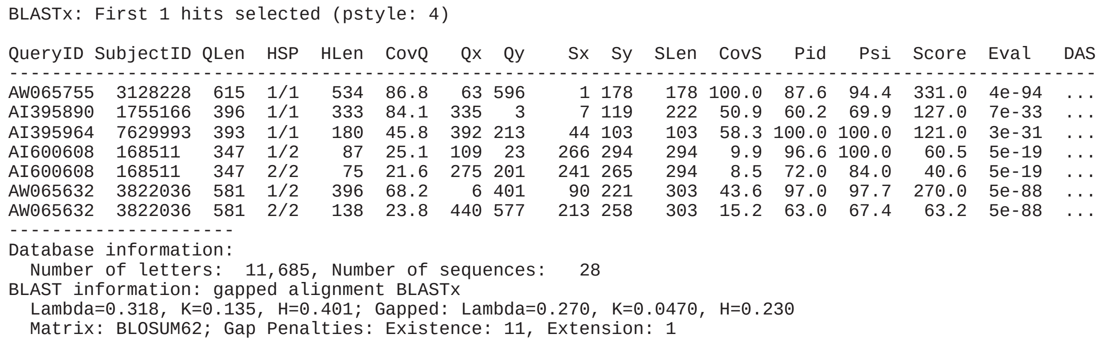
Primary input to MuSeqBox is the output of an NCBI BLAST run. For each BLAST hit, the program derives and saves the following features (Figure 1): query sequence identifier; subject (matching database sequence) identifier; query sequence length (*Qlen*); number of High-scoring Segment Pairs (HSPs); HSP length (*Hlen*); query coordinates of the HSP (from, to); subject coordinates of the HSP; subject sequence length (*Slen*); percent query coverage ($CovQ = 100 \cdot Hlen/Qlen$); percent subject coverage ($CovS = 100 \cdot Hlen/Slen$); percent identical residues in the HSP (*Pid*); percent similar residues in the HSP (*Psi*); number of gap symbols in the HSP alignment (*Ngap*); reading *Frame*; score of the HSP (*Score*); expectation value of the HSP (*Eval*); database identifier, subject description, and subject source (organism).

MuSeqBox APPLICATION TO EST ANNOTATION AND ANALYSIS

EST collections are currently produced for many species as an efficient strategy for gene identification (see http://www.ncbi.nlm.nih.gov/dbEST/index.html). Analysis of the ESTs involves clustering and contig formation and annotation (e.g. Gai *et al.*, 2000). Several filter options of MuSeqBox were particularly designed for EST (and EST contig consensus) sequence annotation and analysis tasks. Figure 1 illustrates such an application with MuSeqBox-processed BLASTX results for five selected maize ESTs.

EST annotation

Tentative annotation of an EST query is implemented by screening for highly significant BLASTX hits against a protein database. In addition to setting the BLASTX parameters at high stringency (low *Eval*), accuracy of annotation by similarity is further improved by two



```
BLASTx: First 1 hits selected (pstyle: 4)

QueryID SubjectID QLen  HSP  HLen  CovQ   Qx  Qy    Sx  Sy  SLen  CovS   Pid   Psi  Score  Eval   DAS
-----------------------------------------------------------------------------------------------------
AW065755  3128228  615  1/1   534  86.8   63 596     1 178   178 100.0  87.6  94.4  331.0  4e-94  ...
AI395890  1755166  396  1/1   333  84.1  335   3     7 119   222  50.9  60.2  69.9  127.0  7e-33  ...
AI395964  7629993  393  1/1   180  45.8  392 213    44 103   103  58.3 100.0 100.0  121.0  3e-31  ...
AI600608  168511   347  1/2    87  25.1  109  23   266 294   294   9.9  96.6 100.0   60.5  5e-19  ...
AI600608  168511   347  2/2    75  21.6  275 201   241 265   294   8.5  72.0  84.0   40.6  5e-19  ...
AW065632  3822036  581  1/2   396  68.2    6 401    90 221   303  43.6  97.0  97.7  270.0  5e-88  ...
AW065632  3822036  581  2/2   138  23.8  440 577   213 258   303  15.2  63.0  67.4   63.2  5e-88  ...
---------------------
Database information:
  Number of letters:  11,685, Number of sequences:   28
BLAST information: gapped alignment BLASTx
  Lambda=0.318, K=0.135, H=0.401; Gapped: Lambda=0.270, K=0.0470, H=0.230
  Matrix: BLOSUM62; Gap Penalties: Existence: 11, Extension: 1
```

Fig. 1. MuSeqBox output table. Column *QueryID* gives the GenBank accessions identifying five maize ESTs that were used as queries in a BLASTX search against the current non-redundant protein database. Features of significant BLASTX hits are tabulated as described in the text. The last three columns (DAS) give the subject sequence description, omitted in the figure for clarity. The descriptions are: 3128228, 60S ribosomal protein L18A [*A.thaliana*]; 1755166, germin-like protein [*A.thaliana*]; 7629993, histone H4-like protein [*A.thaliana*]; 168511, protein cdc2 kinase [*Zea mays*]; 3822036, endo-1,3-1,4-beta-D-glucanase [*Z.mays*]. The number and positioning of the HSPs suggest the following properties: AW065755, full-length coding sequence for L18A; AI395890, 5′-EST; AI395964, 3′-EST; AI600608 and AW065632, alternatively spliced ESTs. Columns NGap and Frame are not displayed (see text).

MuSeqBox parameters that select only those hits with above threshold values for *CovS* and *Pid*. For example, $CovS > 80\%$ and $Pid > 60\%$ would select queries with HSPs that cumulatively extend over more than 80% of the matching protein subject, with overall degree of identity at least 60%. Such criterion would be appropriate for highly conservative annotation. At this level of stringency, the EST may be confidently assumed to encode a homolog of the matching protein subject.

ESTs encoding full-length open reading frames

MuSeqBox allows screening for BLASTX hits of EST queries that match a protein subject with high coverage (e.g. $CovS > 90\%$), including the N- and C-termini. For example, BLASTX results summarized in Figure 1 strongly suggest that the maize EST sequence AW065755 contain the entire open reading frame encoding the maize homolog of the *Arabidopsis thaliana* 60S ribosomal protein L18A. Similar criteria can be set to identify 5′- or 3′-coding ESTs (Figure 1).

ESTs representing potential alternatively spliced transcripts

EST sequences may indicate genes that are potentially alternatively spliced. One such possibility is that the EST has retained an intron that is inefficiently spliced or retained in transcripts expressed in particular tissues. If a matching protein subject is found in a BLASTX search, the BLASTX hit would result in multiple HSPs that are contiguous in the protein subject but separated by an insert in the EST query. An example is provided by the BLASTX hit of the maize EST AI600608 against the maize protein cdc2 kinase (Figure 1). Refined analysis with the spliced alignment algorithm GeneSeqer (Usuka and Brendel, 2000) reveals high-scoring splice sites flanking the insert, strongly supporting its intron status (data not shown). Similar selection criteria are implemented to search for examples of exon skipping.

Other features of the program are described on the program web site and in the software documentation.

ACKNOWLEDGEMENTS

This work was supported in part by NSF grant DBI-9872657. The authors are grateful to Hewlett-Packard Inc. for the gift of two powerful UNIX workstations on which the code was developed.

REFERENCES

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. http://www.ncbi.nlm.nih.gov/BLAST/

Brenner,S.E. (1999) Errors in genome annotation. *Trends Genet.*, **15**, 132–133.

Gai,X., Lal,S., Xing,L., Brendel,V. and Walbot,V. (2000) Gene discovery using the maize genome database ZmDB. *Nucleic Acids Res.*, **28**, 94–96. http://zmdb.iastate.edu/

Koonin,E.V., Aravind,L. and Kondrashov,A.S. (2000) The impact of comparative genomics on our understanding of evolution. *Cell*, **101**, 573–576.

Sonnhammer,E.L. and Durbin,R. (1994) A workbench for large-scale sequence homology analysis. *Comput. Appl. Biosci.*, **10**, 301–307.

Usuka,J. and Brendel,V. (2000) Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. *J. Mol. Biol.*, **297**, 1075–1085.