BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 12 2004, pages 1968–1970
doi:10.1093/bioinformatics/bth172

BLAST2GENE: a comprehensive conversion of BLAST output into independent genes and gene fragments

Mikita Suyama[†], David Torrents[†] and Peer Bork*

Biocomputing, European Molecular Biology Laboratory, Meyerhofstrasse 1,
D-69012 Heidelberg, Germany

Received on January 5, 2004; revised and accepted on January 29, 2004
Advance Access publication March 22, 2004

ABSTRACT

Summary: BLAST2GENE is a program that allows a detailed analysis of genomic regions containing completely or partially duplicated genes. From a BLAST (or BL2SEQ) comparison of a protein or nucleotide query sequence with any genomic region of interest, BLAST2GENE processes all high scoring pairwise alignments (HSPs) and provides the disposition of all independent copies along the genomic fragment. The results are provided in text and PostScript formats to allow an automatic and visual evaluation of the respective region.

Availability: The program is available upon request from the authors. A web server of BLAST2GENE is maintained at http://www.bork.embl.de/blast2gene

Contact: bork@embl.de

INTRODUCTION

BLAST (Altschul *et al*., 1997) is the most widely used program in similarity searches mostly because of its speed and sensitivity. While the use of BLAST has been traditionally restricted to the identification of cDNAs or proteins similar to a query sequence, in recent years it has extended to searches of similarity through large genomic sequences in order to identify new regions of interest (normally genes) or even to solve gene structures. The distribution of genomic segments that can be identified by similarity searches often is highly complex, particularly in regions where numerous local duplications and/or rearrangements occurred (e.g. the TCR region in the human chromosome 7). Consequently, the BLAST (or BL2SEQ; Tatusova and Madden, 1999) results obtained from the comparison of a cDNA or protein query sequence with the genome becomes difficult to interpret. Although several programs for parsing and reformatting the outputs of BLAST have been developed (e.g. Xing and Brendel, 2001; Zhang, 2003), these cannot reliably interpret BLAST results of regions with multiple copies of genes, i.e. with numerous alignments including the same regions in the query.

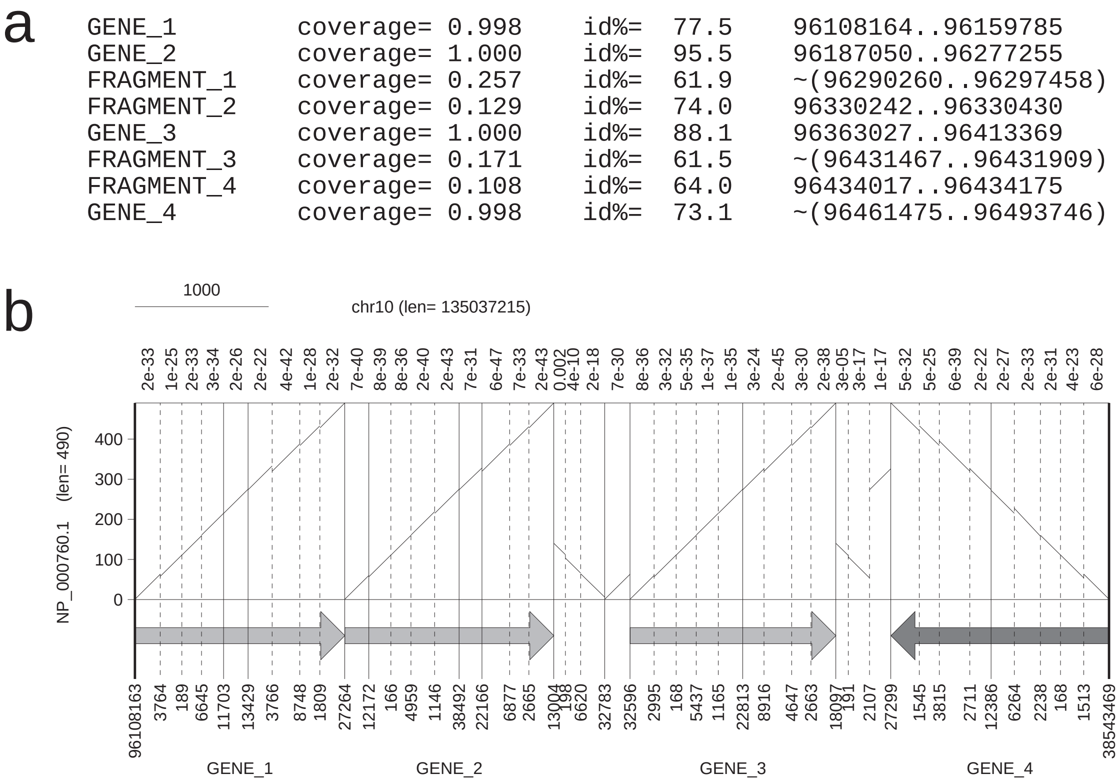
Here we describe the program BLAST2GENE, which identifies all independent partial and complete copies of genes with similarity to a certain genomic region. The program converts the results into PostScript format that allows an easy and fast visual inspection of the region of study.

METHODS AND IMPLEMENTATION

The input to BLAST2GENE is an output of a BLAST search. The action of the program can be divided into three main steps: (1) parsing the BLAST output; (2) identification and delimitation of genes or gene copies based on the parsed output and (3) generation of graphical output. The parser, which is similar to the MuSeqBox program (Xing and Brendel, 2001), converts the information about each high scoring pairwise alignment (HSP) to one text line. In the second step, all these HSPs are first sorted by their position in the genomic sequence. Then, from the upstream HSP, the program examines and connects one of the five downstream HSPs that maintain consistency with the query sequence, i.e. in the same direction and implying fewer and shorter gaps in the query sequence. During the connection of the HSPs we permit two contiguous HSPs to overlap up to 30 residues, since BLAST often extends the alignments beyond the real boundaries to maximize coverage. In order to diminish this effect, we recommend that the matrix option be set as ‘-M PAM70’ with the gap opening penalty ‘-G 10’ and the gap extension penalty ‘-E 1’ in the BLAST option setting (see BLAST tutorial, http://www.ncbi.nlm.nih.gov/BLAST). In contrast, inconsistent HSPs within a possible gene are skipped on the basis of the difference in E-values compared with those of the neighboring HSPs, or because they overlap with another HSP which is more reliable, i.e. with a lower E-value. The entire process is terminated if there are no HSPs that can be connected to the already connected ones. By default, the connected HSPs are considered a complete gene if these HSPs cover >70% of the

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.



```
GENE_1        coverage= 0.998    id%=  77.5    96108164..96159785
GENE_2        coverage= 1.000    id%=  95.5    96187050..96277255
FRAGMENT_1    coverage= 0.257    id%=  61.9    ~(96290260..96297458)
FRAGMENT_2    coverage= 0.129    id%=  74.0    96330242..96330430
GENE_3        coverage= 1.000    id%=  88.1    96363027..96413369
FRAGMENT_3    coverage= 0.171    id%=  61.5    ~(96431467..96431909)
FRAGMENT_4    coverage= 0.108    id%=  64.0    96434017..96434175
GENE_4        coverage= 0.998    id%=  73.1    ~(96461475..96493746)
```

Fig. 1. Text and graphical output of BLAST2GENE. The query and the genomic sequences are human P450 protein subfamily 2C (RefSeq accession no. NP_000760.1) and the human chromosome 10 subsequence from 95 to 100 Mb, respectively. The BLAST options used are -M PAM70 -G 10 -E 1 -F F -e 0.01. (**a**) Text output. There are several formats in the text output depending on the program settings. In this example, coverage and percentage identity to the query sequence and the range on the genomic sequence are listed. (**b**) Graphical output. The query and the genomic sequences are in vertical and horizontal axes, respectively. Each oblique line delimited by the vertical lines represents an HSP. E-values for each HSP are shown on the top of the graph. In the horizontal axis, the regions of genomic sequence without any similarity to the query are not shown for easy interpretation and the length of the excluded genomic regions are indicated in the numbers under the vertical lines. There are three types of vertical lines, which denote genomic distances: thick line (between HSPs >50 kb apart), thin line (between HSPs from 10 to 50 kb apart) and the dotted line (between HSPs <10 kb apart). By defining the minimum coverage in 70%, four genes are represented at the bottom by arrows (light gray, a gene on the sense strand; dark gray, a gene on the complementary strand). The scale for the genomic sequence is shown in the upper left.

query, although this parameter can be tuned by the user. In the third step, the identified HSP belonging to genes and/or gene fragments are plotted along the genomic sequence using the query sequence as a reference.

An example is shown in Figure 1, in which human P450 protein subfamily 2C (RefSeq accession no. NP_000760.1) and the human chromosome 10 subsequence from 95 to 100 Mb are used as the query and genomic sequences, respectively. The BLAST2GENE results reveal, on that chromosomal region, the presence of a gene cluster, which is composed of four full-length genes and additional gene fragments. The graphical output shows the direction of genes and the distances between them.

The program is written in Perl, which takes an output of either BLAST (Altschul *et al*., 1997) or BL2SEQ (Tatusova and Madden, 1999), as input. The output formats are text [Fig. 1(a)] and PostScript [Fig. 1(b)]. It takes 0.56 s to obtain the results shown in Figure 1, which contains 43 HSPs, using a 550 MHz Pentium-III with the Linux operating system.

In summary, BLAST2GENE is capable of interpreting a series of HSPs from a BLAST comparison of a query sequence (normally protein or cDNA) with a genomic region, providing the distribution of all complete or fragmented copies of genes. The two different output formats (text and PostScript) make easy an automatic and large-scale evaluation of the results as well as their visual inspection. BLAST2GENE was successfully applied in the detection of pseudogenes in locally duplicated genes and gene clusters (Hillier *et al*., 2003; Torrents *et al*., 2003). BLAST2GENE can be combined with other programs, such as GeneWise (Birney and Durbin, 1997), in order to obtain a reliable prediction of the exon–intron boundaries of genes. It has the potential to identify many local duplications that as yet have escaped annotation.

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 sequence database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Birney,E. and Durbin,R. (1997) Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 56–64.

Hillier,L.W., Fulton,R.S., Fulton,L.A., Graves,T.A., Pepin,K.H., Wagner-McPherson,C., Layman,D., Maas,J., Jaeger,S., Walker,R. *et al*. (2003) The DNA sequence of human chromosome 7. *Nature*, **424**, 157–164.

Tatusova,T.A. and Madden,T.L. (1999) BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.*, **174**, 247–250.

Torrents,D., Suyama,M., Zdobnov,E. and Bork,P. (2003) A genome-wide survey of human pseudogenes. *Genome Res.*, **13**, 2559–2567.

Xing,L. and Brendel,V. (2001) Multi-query sequence BLAST output examination with MuSeqBox. *Bioinformatics*, **17**, 744–745.

Zhang,H. (2003) Alignment of BLAST high-scoring segment pairs based on the longest increasing subsequence algorithm. *Bioinformatics*, **19**, 1391–1396.