Sequence analysis

NOBLAST and JAMBLAST: New Options for BLAST and a Java Application Manager for BLAST results

Jacques Lagnel1,*, Costas S. Tsigenopoulos1 and Ioannis Iliopoulos2,∗

1Institute of Marine Biology and Genetics, Hellenic Centre for Marine Research, Heraklion 71003 and
2Division of Medical Sciences, University of Crete Medical School, Heraklion 71110, Crete, Greece

Received on October 7, 2008; revised on January 26, 2009; accepted on January 27, 2009
Advance Access publication January 29, 2009
Associate editor: John Quackenbush

ABSTRACT
Summary: NOBLAST (New Options for BLAST) is an open source program that provides a new user-friendly tabular output format for various NCBI BLAST programs (Blastn, Blastp, Blastx, Tblastn, Tblastx, Mega BLAST and Psi BLAST) without any use of a parser and provides E-value correction in case of use of segmented BLAST database. JAMBLAST using the NOBLAST output allows the user to manage, view and filter the BLAST hits using a number of selection criteria.


Contact: lagnel@her.hcmr.gr; ioannis@iti.gr

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
Sequence similarity detection methods are used to predict structural and functional properties of genes and proteins and to detect evolutionary relationships among them. With over 800 completed genome-sequencing projects and many ongoing ones (Bernal et al., 2003), similarity searches are fundamental in order to exploit efficiently all this huge amount of data.

Over the past 30 years, researchers have developed a large panel of powerful methods for detecting gene and protein sequence similarities. Tools have been developed not only to compare one query sequence against a database, but also to compare genomes against other genomes. One of the most commonly used similarity search program is BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990; Altschul et al., 1997). Usually, in order to manipulate the standard NCBI BLAST output someone could not avoid the use of a parser (Catanho et al., 2006; Paquola et al., 2003) and/or programming skills are required (Mangalam, 2002; Paquola et al., 2003). Here, we present NOBLAST, a patch for the NCBI toolkit that provides a new extended tabular output format for various NCBI BLAST programs (Blastn, Blastp, Blastx, Tblastn, Tblastx, Mega BLAST and PSI BLAST). This format can be directly analysed in any spreadsheet and/or any database engine (e.g. MySQL). NOBLAST offers an option to correct the statistics (E-value correction) when split databases are used (e.g. distributed BLAST in a grid or in a cluster) or when BLAST normalization is required (e.g. BLAST with a query set against several databases of different sizes). We also present an additional new viewer/filter for NOBLAST output called JAMBLAST.

2 NOBLAST
2.1 Output option
The new output format is produced using the newly introduced –m 18 and –m 19 options in the patched blastall, blastcl3 and blastpgp command line. Both options produce an extended text tabular delimited file format with all the BLAST hits and their respective attributes. The only difference between the two options is that –m 18 stores each alignment, whereas –m 19 does not. All features of standard NCBI BLAST output are present and additionally some other such as: query and subject index, query and subject length coverage of the alignment, separated query and subject number of gaps, hits rank for a given query and high scoring pairs (HSPs) rank for a given hit (for more details see Supplementary Material). These indexes might be helpful for reassembling BLAST results when split query/databases are used.

2.2 E-value correction method
In order to calculate the E-value, one of the BLAST requirements is the total size of the BLAST database (n) and the number of sequences (N) (Korf, 2003)

\[ E = K \cdot m' \cdot n' \cdot e^{-\lambda \cdot S} \]

\( m' \cdot n' \) is the effective search space, \( m' = m - l \) and \( n' = n - N \cdot l \), where m is the number of letters in the query, K is a minor constant, S is the “raw” score and l is the adjustment to the lengths of the query and database sequences that are used to compensate for edge effects.

To calculate l, n and N are again taken into account:

\[ f(l) = \beta + (\alpha / \lambda) \cdot (\ln(K) + \ln((m - l) \cdot (n - N \cdot l))) \]

and

\[ K(m - l) \cdot (n - N \cdot l) \geq \max(m, n) \]

By default, BLAST uses the $n$ and $N$ of the given database. However, in the case of split databases, E-values should be calculated by taking into account the $n$ and $N$ of the original non-segmented database. Therefore, we introduce a new option -x (for blastall and blastcl3 application) that allows the user to set the number of sequences in the original database. The -x option should be used with the already existed -z option (effective length of database) and the combination of both provides the E-value correction (after parameters -x and -z are set, BLAST uses them each time it calculates $(m'\cdot n')$ and $l$).

Thus, NOBLAST provides exact E-value statistics like mpiBLAST (which, however, requires a Beowulf cluster and does not work on distributed/grid computer) in contrast to other implementations of parallel BLAST like NBlast, dBlast, parallelblast and Blast.pm that provide approximate E-value statistics (Mirto et al., 2008).

For the formatdb application a new -D (Trusted seq NCBI FASTA definition ID) argument is introduced. When an index is created (-o option) the user has now the choice to use it either to parse the unique identifier from a well-formatted NCBI FASTA definition line as a sequence ID or to parse the index sequences of the subject file when a custom BLAST database is used (the limit of 32,767 sequences is now fixed to 2,147,483,647).

NOBLAST is a patch for the NCBI toolkit that produces modified blastall, blastcl3, blastpgp and formatdb executables. We provide the source code as well as pre-compiled executables for some common Windows and Linux platforms.

3 JAMBLAST

3.1 Visualization of the output

Although, the output produced with the –m18/19 option can be directly loaded and analyzed in any spreadsheet or database engine, we provide an additional Java application called JAMBLAST which can be used optionally for the management of the BLAST outputs produced by NOBLAST, using MySQL. It also offers visualization, sorting and filtering of the results according to user-defined criteria (Fig. 1).

JAMBLAST is written in Java and we have tested its functionality under Windows and Linux. It needs a MySQL (MySQL version 4.1 or higher) server and Java (Sun Java jre 1.5.0 or newer) to be present. Similar tools, such as NuclearBLAST (Dienert et al., 2005) and BlastQuest (Farmerie et al., 2005), require web server installation on Linux platform. On the contrary, JAMBLAST runs directly on any personal computer without any special user computer knowledge. Moreover, with JAMBLAST the user can have access to the data, either locally or through network connection (distant MySQL server).

ACKNOWLEDGEMENTS

The authors would like to thank Christos Ouzounis (King’s College, London), Erick Desmarais (University of Montpellier II) and Anastasia Tsagkarakou (N.AGRE.F /Plant Protection Institute, Heraklion) for critical reading of the manuscript, helpful suggestions and comments.

Conflict of Interest: none declared.

REFERENCES


