Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229
GENOMICS PROTEOMICS & BIOINFORMATICS

Application Note

BiDiBlast: Comparative Genomics Pipeline for the PC

João M.G.C.F. de Almeida*

Centro de Recursos Microbiológicos (CREM), Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516 Caparica, Portugal.

Genomics Proteomics Bioinformatics 2010 Jun; 8(2): 135-138. DOI: 10.1016/S1672-0229(10)60015-0

Abstract

Bi-directional BLAST is a simple approach to detect, annotate, and analyze candidate orthologous or paralogous sequences in a single go. This procedure is usually confined to the realm of customized Perl scripts, usually tuned for UNIX-like environments. Porting those scripts to other operating systems involves refactoring them, and also the installation of the Perl programming environment with the required libraries. To overcome these limitations, a data pipeline was implemented in Java. This application submits two batches of sequences to local versions of the NCBI BLAST tool, manages result lists, and refines both bi-directional and simple hits. GO Slim terms are attached to hits, several statistics are derived, and molecular evolution rates are estimated through PAML. The results are written to a set of delimited text tables intended for further analysis. The provided graphic user interface allows a friendly interaction with this application, which is documented and available to download at http://moodle.fct.unl.pt/course/view.php?id=2079 or https://sourceforge.net/projects/bidiblast/ under the GNU GPL license.

Key words: comparative genomics, molecular evolution, sequence batch processing, annotation transfer

Introduction

The increasing availability of the complete sequence for a growing number of genomes has been revolutionizing the way biologists work. The full exploitation of this data bonanza is hampered by the limitations in sequence annotation. These limitations result from an imbalance between the accumulation rate of new sequences, and the throughput of the so called wet bench researchers. The gap is usually filled by *in silico* analysis, mostly done automatically through software pipelines (*e.g.* EMBL Bank to TrEMBL). The results are more often than not stored in secondary databases after a brief assessment due to natural limitations in staff (*1*). This state of affairs results in the need to enforce a most strict set of parameters for the automatic annotation in order to avoid or limit the emergence of artifacts (*e.g.* annotation transfer from analogs).

Today most genome centred databases offer pre-computed comparative genomic results to speed the analysis required by the ordinary user. Those results are often available through rich graphic user interfaces that ease the burden of finding the intended data, although this easiness of usage is counterbalanced by the conservative nature of released datasets. So from time to time a user can stumble upon missing homolog data, especially when

*Corresponding author.
E-mail: jmfa@fct.unl.pt

dealing with genomes with low coverage ratio.

These gaps in the available annotation could be easily plugged by giving to the user the ability to recheck the data with less strict parameters. Bi-directional BLAST is a traditional procedure for the first approach to homolog detection, but nowadays it must be supplemented with domain architecture information, and should be also corroborated by synteny analysis. But domain detection is still a very heavy task in terms of computation time, and synteny is seldom available, so there is room for simpler intermediate solutions, like checking global alignments of both the open reading frame (ORF) and its conceptual product. The extension of the alignment and the similarity can be used to discard BLAST false hits (*e.g.* partial matches due to a shared conserved domain).

Resource Description

Capabilities

BiDiBlast enables the average user with an ordinary personal computer to perform the matching of two sets of sequences by means of a bi-directional (*2-4*) BLAST (*5*) or TBLASTX search. The resulting of local alignment matches can be complemented by a subsequent global alignment in order to evaluate each hit properly. If one of the two sets of sequences happens to be well documented, this application may be used to transfer annotation among the putative orthologs found. It will also allow for the assignment of gene ontology annotation to either bi-directional or simple matches in order to allow assessment of the distribution of ontologically related sequences. Finally for every match the evolution rates dN and dS are determined along with their estimated standard error (*6*).

All these capabilities do not usually concur in a single application, and in BiDiBlast they let the user perform a range of analysis from the customised comparative genomics to molecular evolution measurements on batches of sequences, or to the equivalent to a more detailed BLAST search against a limited set of sequences.

Implementation

The pipeline code is entirely written in JavaSE 1.5 (Sun Microsystems, Inc.), using the DB4O (Versant Corp.) to manage the data and the results. The BioJava library (*7*) is used solely for translation work and translation table management. The searches for local similarity among sequences are done by a local installation of the NCBI BLAST tool (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/). Global alignment of ORF sequences is done by the tool 'align0' (*8*) from the FASTA package. This tool is special because it does not penalise the introduction of gaps at the ends of the smaller sequence, thus allowing for a smoother result. 'Stretcher' from the EMBOSS suite (*9*) was used to align pair of translated ORF because the available 'align'/'align0' binaries are irresponsive to substitution matrix selection. This renders them ineffective to align carefully protein sequences in Windows environments. The tool creator acknowledged the problem, but he didn't have the opportunity to solve this problem in time. Molecular evolution rates are also calculated through the yn00 tool from PAML package (*10*). This set of tools may be updated separately from the pipeline application by the users themselves. They also limit the portability of this application to environments other than 32 bits of MS Windows to the availability of native binaries for those tools. This portability will also require that the Java code should be changed where direct calls are made to the operating system. Although compiled for JavaSE 1.5, the application runs fine also on JavaSE 1.6.

The input for the pipeline consists of two text files containing sequences in FASTA format, one is the query set, and the other plays the role of reference (**Figure 1**). These sequences should be uninterrupted ORFs if the analyses are to be carried in full, otherwise only the bi-directional BLAST procedure would make sense. The user may order the application to compile the sequences as BLAST binary databases before invoking the Blast procedure or may supply the databases already compiled. Other text files may be supplied to enrich the analysis, namely a GO Slim terms (*11*) list (http://downloads. yeastgenome.org/literture_curation/go_terms.tab), and a map assigning GO Slim terms to the relevant reference sequences (http://downloads.yeastgenome.org/literature_ curation/

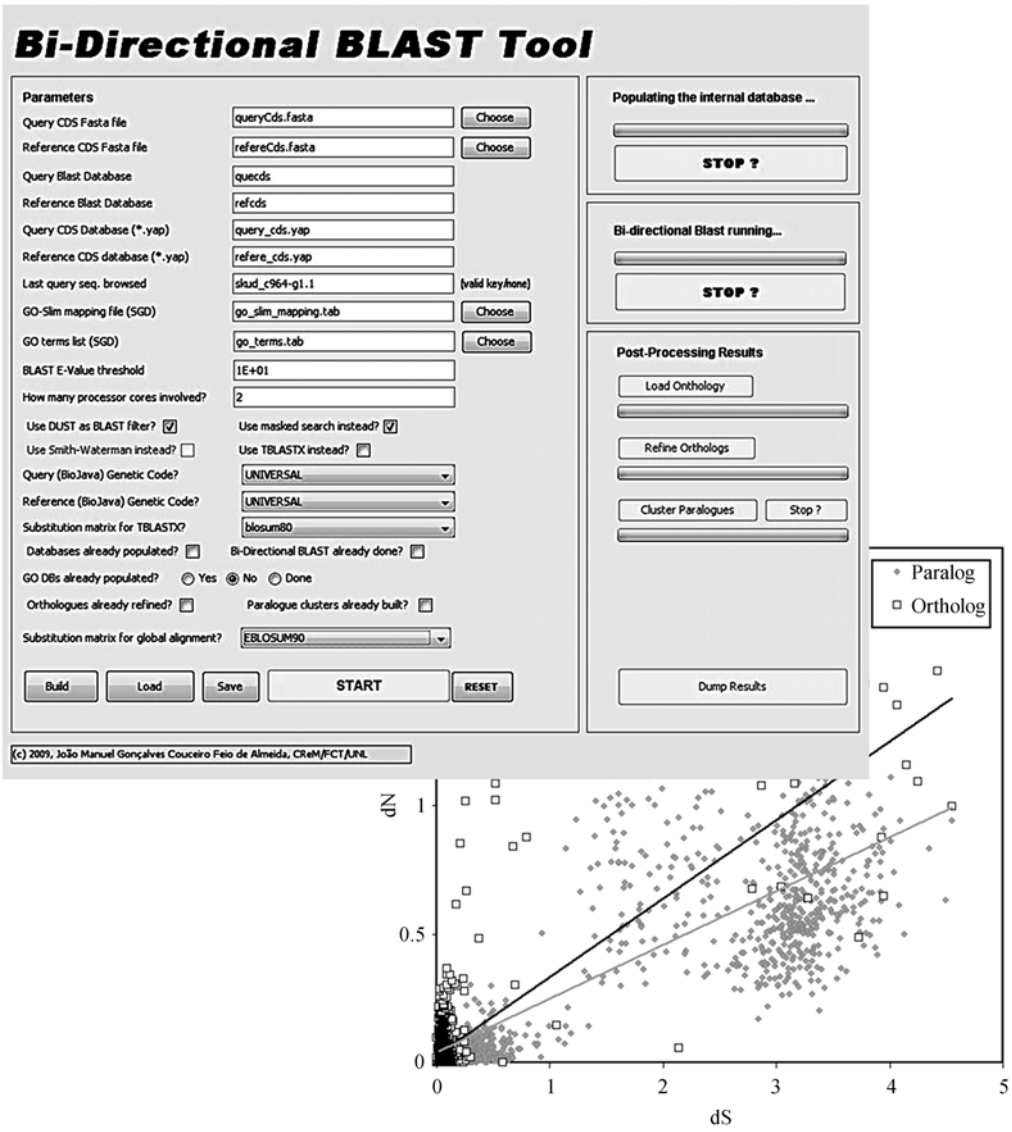

Figure 1 BiDiBlast graphic user interface after starting the application, and an example of an analysis enabled by this software showing calculation of average values for dN/dS rates for the putative orthologous and paralogous ORFs.

go_slim_mapping.tab) as available from *Saccharomyces* Genome Database project (ftp://ftp.yeastgenome.org/yeast/). Similar files for other genomes may be supplied instead, provided the column arrangement is respected.

The first stage of the program execution is to upload the sequences into the internal database. Then the bi-directional BLAST procedure begins matching up every sequence in the query set against the reference sequences. At each step the top hit in the results list is subjected to the reverse process. If the top hit in this final search is the starting sequence, a bi-directional hit is scored, and uni-directional hits are also recorded as they may point to paralogous ORFs. If the sequences that are being compared are not from close related taxa, TBLASTX variant should be the option of choice because matches are evaluated at the conceptual translation level. In this case the local alignments will be shorter and ungapped, therefore less emphasis should be put in the resulting BLAST statistics. The procedure will also require the selection of an adequate substitution matrix, and it will take longer to complete. The user is also given the possibility of using DUST filter and masking.

In a second stage, GO Slim data are imported into the internal database and mapped only to BLAST hits,

through the reference sequences to the ones being queried. The results from the previous BLAST search are now complemented by two rounds of paired global alignment. The first aligns the original nucleotide sequences, while the second aligns the conceptual translation products, and concomitantly a codon-wise alignment for the former sequences is generated. For this alignment procedure to be carried with best results, the substitution matrix should be chosen by the user according to the estimate of the average similarity between both proteomes. The results from this last round are most important when the matched sequences are more divergent, and TBLASTX is used. The translation of the sequences is performed according to a particular genetic code. This is specified for each set of sequences, and is also applied to the eventual TBLASTX searches.

The final stage consists of calculations done over the product and codon wise alignments. Several statistics are calculated about observed substitution frequency against base position in the codon, and the degree of conservation among amino acids. Evolution rates are calculated as well.

Results and Discussion

Each application run may take up to three days for ORFomes of the size of those known for *Sacharomyces* species in a low end PC computer. The total data and results can be browsed with the proprietary DB4O Object Browser (Versant Corp; http://developer.db4o.com/files/folders/objectmanager_1746/entry24858. aspx), but they are mainly intended to be imported as text delimited files into a spreadsheet editor or relational database management system (*e.g.* MS Access or mySQL). Such applications will allow the user to filter, process, and explore the results in the most efficient way.

The approach taken in the inception of this application was to minimize the amount of filtration of input data (*e.g.* malformed ORFs with intervening stop codons) or the yielded results, in order to give the user a greater flexibility at analysis time. This empowerment should compel the user to be more careful in the evaluation of the result dataset. Artifacts are bound to arise and they must be discarded before undertaking any kind of analysis. The most straightforward filter to apply when looking for putative ortholog sequences should be the paired extension of the global alignment relative to the query sequence length. This simple operation will allow for the detection of one of the most common sources of false hits, partial alignments due to a shared conserved domain.

This is an evolving application with new versions already in development.

Acknowledgements

The author wishes to thank all the researchers, institutions, and enterprises that made their code or tools freely available to non-profit projects like this.

References

1 Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28: 45-48.
2 Rivera, M.C., *et al.* 1998. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. USA* 95: 6239-6244.
3 Bork, P., *et al.* 1998. Predicting function: from genes to genomes and back. *J. Mol. Biol.* 283: 707-725.
4 Tatusov, R.L., *et al.* 1997. A genomic perspective on protein families. *Science* 278: 631-637.
5 Altschul, S.F., *et al.* 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.
6 Yang, Z. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17: 32-43.
7 Holland, R.C., *et al.* 2008. BioJava: an open-source framework for bioinformatics. *Bioinformatics* 24: 2096-2097.
8 Myers, E.W. and Miller, W. 1988. Optimal alignments in linear space. *Comput. Appl. Biosci.* 4: 11-17.
9 Rice, P., *et al.* 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* 16: 276-277.
10 Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24: 1586-1591.
11 Camon, E., *et al.* 2003. The Gene Ontology Annotation (GOA) Project-Application of GO in SWISS- PROT, TrEMBL and InterPro. *Comp. Funct. Genomics* 4: 71-74.