BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 4 2001
Pages 385–387

PhyloBLAST: facilitating phylogenetic analysis of BLAST results

Fiona S. L. Brinkman[1], Ivan Wan[2], Robert E. W. Hancock[1], Ann M. Rose[3] and Steven J. Jones[2,]*

[1]Department of Microbiology and Immunology, University of British Columbia, Vancouver, Canada V6T 1Z3, [2]Genome Sequence Centre, BC Cancer Agency, Vancouver, Canada V5Z 4E6 and [3]Department of Medical Genetics, University of British Columbia, Vancouver, Canada V6H 3N1

Received on September 8, 2000; revised on December 11, 2000; accepted on December 13, 2000

ABSTRACT

Summary: PhyloBLAST is an internet-accessed application based on CGI/Perl programming that compares a users protein sequence to a SwissProt/TREMBL database using BLAST2 and then allows phylogenetic analyses to be performed on selected sequences from the BLAST output. Flexible features such as ability to input your own multiple sequence alignment and use PHYLIP program options provide additional web-based phylogenetic analysis functionality beyond the analysis of a BLAST result.

Availability: This program is available from http://www.pathogenomics.bc.ca/phyloBLAST/ and the source code is freely available from the authors.

Contact: info@pathogenomics.bc.ca

As the number of genomes sequenced increases, phylogenetic analysis of sequence data is becoming a more useful, and often essential, tool. Phylogenetic analysis is useful not only for the derivation of organismal phylogenies, but also for delineating orthologous and paralogous relationships between the large families of homologous sequences being discovered through genome sequencing. Determining such evolutionary relationships has many uses, including facilitating predictions about protein function (Paulsen *et al.*, 1998) and identifying horizontal gene transfer events, however it is often a time consuming process. We initially developed PhyloBLAST to aid us in more quickly identifying potential horizontal gene transfer events between bacteria and eukaryotes, however, we realized that PhyloBLAST could be of broad utility for phylogenetic analysis of any proteins, in particular identification of orthology versus parology, and so we adapted it for public use.

In short, PhyloBLAST is a web-based application that compares a given protein sequence to a protein database using WU-BLAST2 (Gish, unpublished; Altschul *et al.*, 1990) and allows the user to select sequences in the resulting BLAST output (or other sequences they supply) for further phylogenetic analysis. Phylogenetic analyses are performed using a combination of ClustalW for multiple sequence alignments and the PHYLIP 3.5c package of programs for phylogenetic tree building (http://www.ibb.waw.pl/docs/PHYLIPdoc/; Felsenstein, 1989; Thompson *et al.*, 1994). PhyloBLAST contains a number of features useful for studying evolutionary relationships of proteins and allowing further application flexibility, including:

— Organism and gene information is added to the BLAST output and to the phylogenetics trees subsequently generated.

— After the initial BLAST analysis is performed, the user may select sequences (in the list of BLAST hits) for further analysis, simply by clicking boxes next to each sequence of interest. The user can select either full-length sequences, or the segments of sequences found similar by BLAST (the latter is referred to as 'HSP segment only' for 'high scoring pair segment only'). Full-length sequences can provide more informative sites for the alignment, however due to the domain-nature of proteins in some cases the use of segment pairs may be more appropriate to facilitate alignment of only homologous protein domains.

— The user can also paste in a box, at the end of the BLAST output, additional sequences they wish to use for the analysis.

— Phylogenetic trees may be calculated using two very differing methods from the PHYLIP package programs, with or without bootstrapping.

— The user may also view an alignment of the sequences selected, or can enter in their own multiple sequence

*To whom correspondence should be addressed.

alignment file for performing phylogenetic analyses using an edited alignment (in standard PHYLIP interleaved format).

— Full options are available for the phylogenetic analyses.

— Phylogenetic analyses may be e-mailed or the user notified by e-mail when the analysis is done. This is particularly useful when the users wish to save their results or when bootstrapped trees are computed, due to their long computation times.

— The user may obtain phylogenetic trees as either a generic treefile, as a graphic, or as an ASCII text-based tree graphic that contains hyperlinks to further information about the sequences.

These features result in a versatile application that enables the user to also generate phylogenetic trees using other sequences/alignments they supply, essentially bypassing the BLAST portion of the program, and using it instead as a web-interface for PHYLIP. Previously, web-interfaces for PHYLIP have been developed (http://www.sdmc.krdl.org.sg:8080/~lxzhang/phylip/; http://www.bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html; Lim and Zhang, 1999), however they do not merge the PHYLIP programs together for tree construction as PhyloBLAST does, and they do not integrate a protein database and BLAST analysis capability into the application. The Bork Group's BLAST2 & Orthologue Search (http://www.Bork.EMBL-Heidelberg.DE/Blast2e/; Yuan *et al.*, 1998) does perform phylogenetic analyses based on a BLAST result, however the phylogenetic analyses are performed in an automated fashion for the specific purpose of identifying orthologs. While these other tools have their own utility, PhyloBLAST is suitable for those who wish to have the convenience of access to selected PHYLIP programs in a web-based format that links the programs together, and incorporates BLAST analysis with fully functional PHYLIP programs.

Use of the program is simple: at the opening webpage of PhyloBLAST the user pastes into a box a protein sequence which is submitted to BLAST analysis to compare the sequence to an in-house database comprising SwissProt/TREMBL sequences derived from a locally maintained SRS database (Etzold *et al.*, 1996). Options on the opening page allow the user to manipulate common BLAST settings, as well as features added to the BLAST output, including how many pairwise phylogenetic distances are to be calculated (described further below). A help file is provided on-line to aid the user with understanding settings.

The resulting initial output contains buttons at the top used for selecting further analyses, followed by an enhanced BLAST output that contains additional information not available from WU-BLAST2. These enhancements include a graphical display of the BLAST results developed from a perl module by Alessandro Guffanti (http://www.hercules.tigem.it/Biomodules.html), descriptions of the organism name and gene name associated with each hit sequence, a letter code indicating whether the hit organism is a member of the eukarya/bacteria/archaea (E, B, A), and finally pairwise phylogenetic distances calculated between the user's sequence and each hit sequence are shown.

The latter phylogenetic distances are calculated using PHYLIP PROTDIST for both 'protein segments' identified by BLAST (i.e. the HSPs), and 'full-length proteins' (distances calculated from a ClustalW alignment of the users sequence and the full-length hit sequence). These distances may aid the researcher in identifying which sequence in the BLAST output is most similar to their query sequence, and the lowest distance (corresponding to the most similar sequence, other than itself) is coloured in red for easier identification (it is not necessarily the top sequence listed in the BLAST output). Note that whenever there are two HSPs for a given hit sequence, the distance is calculated for the most significant HSP and the value coloured in green to warn the user that there is more than one HSP. Currently these distances are calculated using default settings, however further phylogenetic analysis allows for a more customized study.

Further analysis may be performed on the BLAST output by selecting the appropriate check boxes next to sequences to be analyzed further, and then selecting a button at the top of the page for the analysis desired. The analyses available include three which do not involve the construction of phylogenetic trees: (1) obtaining a FASTA file of the sequences (useful for further manipulation of the sequences using other programs); (2) obtaining a ClustalW alignment of the sequences (useful for viewing how the sequences you selected align with each other using default settings) and (3) obtaining a distance matrix (based on a ClustalW alignment and PHYLIPs PROTDIST). While ClustalW alignments used are currently produced using default settings, the user is given the option to paste in their own alignment for further phylogenetic analysis, so they may use hand-edited or non-default alignments.

Phylogenetic trees can be constructed, based on the selected sequences, using either the Neighbor-joining or Parsimony methods of the PHYLIP 3.5c package. For those not familiar with phylogenetic analyses, Neighbor-joining represents one of the more popular methods for analysis, in part due to its speed of computation and its use of distances. This method's incorporation of distances into the tree branch lengths is useful for visualizing which sequences are more related to each other. Parsimony uses a very different approach to tree construction that

is included to complement the Neighbor-joining distance matrix method. Note that it does not incorporate distances into the tree branch lengths. For further description of these methods, the PhyloBLAST on-line help file provides a link to PHYLIP program descriptions and reviews (e.g. Saitou, 1996).

For both phylogenetic methods, additional bootstrap analyses may be performed. Bootstrapping provides the user with a statistical measure of the reliability of the branching order in the tree. For example, if performed using 50 replicates (the maximum we will initially allow, due to computational resources required), the resulting tree will indicate on each branch the number of times out of 50 that the branching order shown was observed when 50 trees were calculated based on data that was subjected to subtle random perturbations. This is a well-known measure of the reliability of the tree, though others may be included in the future. Note that bootstrapping does not provide an indication of the relatedness between sequences through branch lengths, as a basic Neighbor-joining analysis does, and so commonly a basic Neighbor-joining tree, along with the bootstrap values for each tree node, are reported together.

The trees produced may be viewed either as a JPEG graphic, or as an ASCII text-based graphic tree. The latter includes hyperlinks from each protein accession number in the tree to further information about each sequence. This is useful when you wish to browse the tree, viewing more information about each sequence in a particular clade. Note that the current Drawtree and Drawgram programs used in PHYLIP for generating JPEG graphics of trees may be replaced in the future with ones more suitable for large trees and more suitable for higher resolution graphics. However, the generic 'treefile' is provided with the ASCII tree output, and this treefile is a standard format that may be imported into all other tree-drawing programs we are aware of.

As alluded to earlier, all options for the PHYLIP programs are available for manipulation. Programs are combined together as needed for each analysis, so the user can perform an integrated, yet customized, phylogenetic analysis in one step. In order to adequately handle the anticipated use of this web-based application by the public, all analysis requests are routed on an equal-share basis to any one of 14 processors in a network at the Genome Sequence Centre, Vancouver, Canada. If load on the server is not a problem, we anticipate that further phylogenetic analysis methods will be added and greater flexibility allowed (e.g. more replicates allowed for bootstrap analysis).

In PhyloBLAST, we attempt to combine the useful properties of BLAST, SwissProt, and PHYLIP into a tool relevant for today's post-genome age. As the number of sequences with homology to a given sequence increases dramatically, it becomes increasingly important to complement one-dimensional BLAST analysis with the two-dimensional view of relationships in a phylogenetic tree, and PhyloBLAST facilitates such analysis.

ACKNOWLEDGEMENTS

This application was developed for the Pathogenomics Project (http://www.pathogenomics.bc.ca/), which is funded by the Peter Wall Institute for Advanced Studies, Vancouver, Canada.

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.

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Meth. Enzymol.*, **266**, 114–128.

Felsenstein,J. (1989) PHYLIP—Phylogeny inference package (Version 3.2). *Cladistics*, **5**, 164–166.

Lim,A. and Zhang,L. (1999) WebPHYLIP: a web interface to PHYLIP. *Bioinformatics*, **15**, 1068–1069.

Paulsen,I.T., Sliwinski,M.K. and Saier,M.H.,Jr (1998) Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *J. Mol. Biol.*, **277**, 573–592.

Saitou,N. (1996) Reconstruction of gene trees from sequence data. *Meth. Enzymol.*, **266**, 427–449.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Yuan,Y.P., Eulenstein,O., Vingron,M. and Bork,P. (1998) Towards detection of orthologues in sequence databases. *Bioinformatics*, **14**, 285–289.