BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 6 2003, pages 782–783
DOI: 10.1093/bioinformatics/btg064

Phylogenetic Web Profiler

Philip Wong[1], Grigory Kolesov[2], Dmitrij Frishman[2] and Walid A. Houry[1,]*

[1]Department of Biochemistry, University of Toronto, Medical Sciences Building, 1 King's College Circle, Toronto, ON M5S 1A8, Canada and [2]Institute for Bioinformatics, Ingolstädter Landstraße 1, GSF—National Research Center for Environment and Health, 85764 Neuherberg, Germany

Received on September 27, 2002; revised on November 25, 2002; accepted on December 10, 2002

ABSTRACT

Summary: Phylogenetic Web Profiler (PWP) is a web-based service designed to perform phylogenetic profiling of proteins against genomes. The current version offers a selection of 63 completed genomes and available plasmids as annotated in the PEDANT genome database. Unlike currently available applications, this tool offers several choices of ortholog prediction parameters including E-value cutoff, percent length difference tolerance, and annotation similarity. Additional features include tight integration with the PEDANT database and tools to analyze properties of predicted proteins. PWP should prove very useful for the analysis of functional-linkage between proteins.

Availability: The PWP server is available at http://pedant.gsf.de/pwp/. The PWP functional predictions are also offered as part of the PEDANT genome analysis server (http://pedant.gsf.de).

Contact: walid.houry@utoronto.ca

With the availability of a relatively large number of fully sequenced genomes, the current challenge in bioinformatics resides in our ability to assign function to newly discovered proteins using computational approaches. One such approach is based on correlated evolution of genes, pathways, and complexes (Gaasterland and Ragan, 1998; Pellegrini *et al.*, 1999). This approach has been termed phylogenetic profiling and has been used to infer functional linkage between proteins. Research into this technique has increased in recent literature underscoring its importance in this functional-genomics era. Because of its high computational cost, however, online application of this technique has mainly taken the form of static databases (Mellor *et al.*, 2002; Liberles *et al.*, 2002) which contain pre-computed profiles based on a single set of ortholog prediction parameters. Here, we provide a publicly available online tool, termed Phylogenetic Web Profiler (PWP), which performs dynamic phylogenetic profiling on 63 completed genomes and available plasmids in which the user is given the option of selecting parameters that form the basis for generating the profiles. Tight integration into the highly annotated PEDANT database (Frishman *et al.*, 2001) and other downstream tools to analyze the properties of proteins of interest are also provided.

*To whom correspondence should be addressed.

Profiler options

Phylogenetic profiling is the process of building profiles which are strings of 1's and 0's corresponding to the existence or absence of protein orthologs across selected genomes. Based on a recent analysis, it has been shown that proteins with similar profiles are more likely to be functionally linked (Pellegrini *et al.*, 1999). Using PWP, the user has the option of profiling against the organism as a whole or against available plasmids and chromosomes separately. To carry out profiling, the user can either input the PEDANT identifier of the protein of interest or paste the amino acid sequence of that protein. Currently, three ortholog determination parameters can be varied by the user.

(1) To account for the non-uniform rate of sequence divergence amongst different ortholog families, the option of PSI-BLAST (Altschul *et al.*, 1997) based E-value cutoff is provided. Stringent cutoffs are expected to eliminate false positives in ortholog prediction of more conserved proteins while more relaxed cutoffs will allow detection of more diverged proteins.

(2) The option to specify tolerances to differences in length between the query and hit proteins is provided. Comparison of protein lengths will improve ortholog prediction for proteins with conserved domains. However, when orthologs are products of fission events, using stringent length cutoffs may produce false negatives in generated profiles. To

help compensate for this phenomenon, the program predicts ortholog fission by searching for adjacent genes coding for non-overlapping regions of the same protein.

(3) Finally, the option of comparing annotations between query and hit proteins by word similarity is available as this may help to detect highly sequence divergent orthologs. The annotations are obtained from the PEDANT database.

There are other parameters that can be varied by the user. Options to restrict profiling to certain NCBI-based evolutionary lineages (Wheeler *et al.*, 2000) are also provided. This is helpful for proteins that mainly coexist under one lineage. For example, functional-linkage between the GroEL chaperone and its cofactor GroES will be more evident if the profiling is restricted to bacterial species since both of these proteins are mainly found in bacteria while most archaea only have the GroEL and not the GroES homologues. Likewise, if a protein of interest has no proteins of similar profile when searching against all organisms, then one can proceed to systematically restrict lineages until such proteins are generated. However, one should keep in mind that the likelihood of functional-linkage between proteins may decrease when shorter profiles are considered. In addition, in order to predict what proteins might act as analogous replacements to the query protein in other organisms, PWP provides the option of searching for hits that have an inverted profile to that of the query protein (Liberles *et al.*, 2002). An inverted profile refers to a profile whereby 1's in the query protein profile are replaced with 0's and vice versa.

The program provides a detailed output in which the profile of the query protein is displayed together with profiles from other proteins. The profiles are sorted by bit difference. From this output page, the user can retrieve sequences in FASTA format and can proceed to examine the properties of the displayed proteins such as molecular weight, isoelectric point, and secondary structure class prediction.

Integration

Genomic data for the program was obtained from completely sequenced genomes stored at PEDANT (Frishman *et al.*, 2001). The web program is integrated into the PEDANT framework via links to functional annotation pages of displayed proteins. Thus, predicted functional linkages can be assessed with annotation from a variety of data sources such as the MIPS functional category predictions. Functional linkages can also be compared with those predicted by SNAPper (Kolesov *et al.*, 2001) which is also available as part of the PEDANT server. Visualization of profiles as NCBI-based phylogenetic trees (Wheeler *et al.*, 2000) and the display of underlying PSI-BLAST data used to generate the profiles is provided. For convenience, proteins are linked to Genbank entries. Links to other tools such as PhylProm (Liberles *et al.*, 2002) and STRING (Snel *et al.*, 2000) are also available.

Methods

An all-against-all forward and reverse complete genome PSI-BLAST (Altschul *et al.*, 1997) search was carried out and stored in PEDANT as MySQL tables. Phylogenetic profiles were pre-computed for available *E*-value, length cutoff, and annotation options and stored in MySQL tables. Profiles restricted to certain lineages are generated online by removal of bits from stored profiles according to lineage cutoff. Subsequent clustering of proteins is done based on the protein submitted and the bit difference between profiles. Programs were written as CGI perl scripts with output in HTML format. Access to MYSQL tables was provided by the DBI module. Graphs are generated using the GD and GDGraph modules available at http://www.cpan.org.

ACKNOWLEDGEMENTS

We thank Jimmy Le, Jennie Yum, and Martin Mokrej for technical support.

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.

Frishman,D., Albermann,K., Hani,J., Heumann,K., Metanomski,A., Zollner,A. and Mewes,H.W. (2001) Functional and structural genomics using PEDANT. *Bioinformatics*, **17**, 44–57.

Gaasterland,T. and Ragan,M.A. (1998) Microbial genescapes, phyletic and functional patterns of orf distribution among prokaryotes. *Microb. Comp. Genomics*, **3**, 199–217.

Kolesov,G., Mewes,H.W. and Frishman,D. (2001) SNAPping up functionally related genes based on context information: a colinearity-free approach. *J. Mol. Biol.*, **311**, 639–656.

Liberles,D.A., Thorén,A., von Heijne,G. and Elofsson,A. (2002) The use of phylogenetic profiles for gene predictions. *Current Genomics*, **3**, 131–137.

Mellor,J.C., Yanai,I., Clodfelter,K.H., Mintseris,J. and DeLisi,C. (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.*, **30**, 306–309.

Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.

Snel,B., Lehmann,G., Bork,P. and Huynen,M.A. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.*, **28**, 3442–3444.

Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L., Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.