BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 2 2006, pages 248–250
doi:10.1093/bioinformatics/bti757

Systems biology

PRODISTIN Web Site: a tool for the functional classification of proteins from interaction networks

Anaïs Baudot[1,†], David Martin[1,4,†], Pierre Mouren[1,†], François Chevenet[3], Alain Guénoche[2], Bernard Jacq[1] and Christine Brun[1,*]

[1]LGPD/IBDM, UMR6545 CNRS and [2]Institut de Mathématiques de Luminy, UPR9016 CNRS, Parc Scientifique et Technologique de Luminy, Case 907, 13288 Marseille cedex 9, France and [3]GEMI, UMR IRD/CNRS9926, IRD - BP 64501, 911 Avenue Agropolis, 34394 Montpellier cedex 5, France

Received on July 28, 2005; revised on October 17, 2005; accepted on October 28, 2005
Advance Access publication November 3, 2005
Associate Editor: John Quackenbush

ABSTRACT

Summary: The PRODISTIN Web Site is a web service allowing users to functionally classify genes/proteins from any type of interaction network. The resulting computation provides a classification tree in which (1) genes/proteins are clustered according to the identity of their interaction partners and (2) functional classes are delineated in the tree using the Biological Process Gene Ontology annotations.

Availabitily: The PRODISTIN Web Site is freely accessible at http://gin.univ-mrs.fr/webdistin

Contact: brun@ibdm.univ-mrs.fr

INTRODUCTION

Protein–protein interaction maps are now available for at least four eukaryotic model organisms: the budding yeast (Ito *et al*., 2001; Uetz *et al*., 2000), the worm (Li *et al*., 2004), the fruit fly (Formstecher *et al*., 2005; Giot *et al*., 2003) and human (Stelzl *et al*., 2005; Rual *et al*., 2005). These maps form large intricate networks leading to a renewed vision of cell biology as an integrated system. However, extracting and revealing the functional information they contain depends on our ability to analyze them in detail. Indeed, although they are far from being complete, the size and the complexity of these networks (∼6000, 5000, 20 000, 5500 interactions/graph edges in yeast, worm, fly and human, respectively) make their functional analysis a difficult task.

In order to extract this information, we developed two years ago a bioinformatic method named PRODISTIN (PROtein DISTance based on INteractions) which allows a functional classification of the proteins according to the identity of their interacting partners (Brun *et al*., 2003). The central idea in this interaction-based functional clustering is to compare interaction partners for all protein pairs, assuming that the more two proteins share interacting partners, the more they should be functionally related. Applying it to the yeast protein–protein interaction network, we have previously shown that the method (1) clusters proteins participating in the same cellular processes in the same functional classes, (2) predicts function for unknown proteins and (3) is statistically valid (Brun *et al*., 2003). In addition, we showed that PRODISTIN can successfully be used to study the evolutionary fates of the yeast duplicated genes (Baudot *et al*., 2004). We now propose an automated version of the PRODISTIN method for the community, through a web interface called Prodistin Web Site (PWS).

SERVER OVERVIEW

This server can be used to functionally classify genes/proteins from interaction networks. The resulting computation provides the user with a classification tree in which the network genes/proteins are clustered according to their functional similarity (Step 1, see below) and the functional classes are identified using the Biological Process Gene Ontology annotations of the genes/proteins (Step 2, see below).

The PWS backbone is mainly written in PHP with calls to Perl scripts and C programs.

Step 1. The user enters the interaction network to be analyzed as a list of binary interactions and specifies a minimal connectivity threshold for genes/proteins to be classified. The use of this threshold is intended to eliminate poorly connected genes/proteins from the classification process for which the functional analysis is likely to be biased by the presence of putative false-positive interactions (default value: 3).

The server then computes the Czekanowski–Dice distance between all possible pairs of proteins (for details, see Brun *et al*., 2003). The obtained values are subsequently clustered using the BioNJ algorithm (Gascuel, 1997), leading to a classification tree (Fig. 1, Step 1).

Step 2. Functional classes (Prodistin classes) are then identified in the tree. They correspond to the largest possible subtrees composed of at least *X* proteins sharing the same Biological Process GO annotations and representing at least *Y%* of the individual class members for which an annotation is available (*X* and *Y* are user defined). For this, the PWS is taking into account not only the GO term(s) annotating the genes/proteins contained in the tree but also the complete hierarchy of parent terms. This hierarchy is retrieved by PWS through GOToolBox, a software suite devoted to gene sets analysis based on associated GO terms that we developed recently

*To whom correspondence should be addressed.
†These authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.
[4]Present address: Centre for Molecular Medicine and Therapeutics, University of British Columbia Vancouver, BC, Canada.

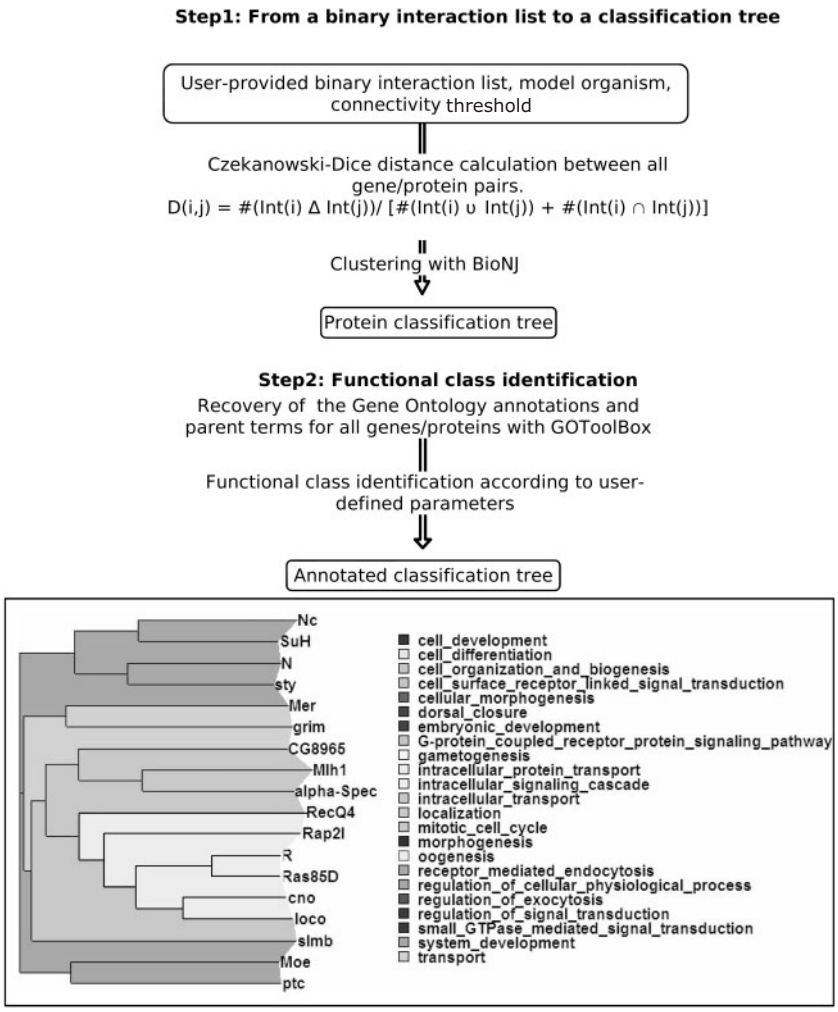

Fig. 1. Synopsis of the Prodistin Web Site and output.

(Martin *et al*., 2004). Once the terms shared by at least *X* genes/proteins in the same subtree are selected, the more precise ones (the deeper in the ontology) are kept to annotate the class. In order to assess the statistical quality of the annotation(s) attributed to the functional classes by the method, a *P*-value of the over-representation of the proposed GO term(s) in the class compared with the tree is calculated using the hypergeometric distribution. This calculation constitutes an optional step of the computation (Step 3).

In metazoan organisms, genes/proteins may be annotated with a very large number of GO terms encompassing a broad range of biological processes. For instance, the *Drosophila* wingless gene is annotated with 48 Biological Process GO terms. Because this situation may blur the network analysis, it may be useful to identify functional classes with only subparts of the Biological Process ontology. For this purpose, the user can select a GO ID corresponding to the GO term to be used as a root term for the subpart of the ontology. The default root value corresponds to the complete Biological Process ontology root term.

The result of the computation is then visualized as a coloured classification tree using an integrated TreeDyn module (http://www.treedyn.org) as a tree viewer (Fig. 1, Step 2). All intermediate files created during the computation are downloadable for further investigations. The provided formats allow the user to load the generated files into an external TreeDyn program for a subsequent deeper analysis of the obtained results.

APPLICATION

The Prodistin Web Site can be used to analyse any kind of biological networks (protein–protein, genetic, protein–DNA). The class annotation process is available for all model organisms supported by GOToolBox (currently *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* and *Saccharomyces cerevisiae*).

CONCLUSION

Prodistin Web Site is a tool allowing the extraction of a part of the functional information contained in interaction networks. Functional classes are identified by a computation of the network and functional annotations are provided with all of them. The combination of these two features makes PWS different from other existing tools. Indeed, for instance, the 'Significant Attribute Plugin' provided with Cytoscape (Shannon *et al*., 2003) is designed to search

for aggregation of attribute values such as annotations in networks without any computation of the network itself, and MCODE (Bader and Hogue, 2003) proposes a computation of the network for the identification of modules without any functional annotations. For these particular reasons, PWS should be valuable to the research community.

ACKNOWLEDGEMENTS

A.B. and D.M. were supported by fellowships from the French 'Ministère de l'Enseignement Supérieur et de la Recherche'. This work was supported by an ACI IMPBio grant (EIDIPP project) to B.J.

Conflict of Interest: none declared.

REFERENCES

Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinfoinformatics*, **4**, 2.

Baudot,A. *et al*. (2004) A scale of functional divergence for yeast duplicated genes revealed from analysis of the protein–protein interaction network. *Genome Biol*., **5**, R76.

Brun,C. *et al*. (2003) Functional classification of proteins for the prediction of cellular function from a protein–protein interaction network. *Genome Biol*., **5**, R6.

Formstecher,E. *et al*. (2005) Protein interaction mapping: a *Drosophila* case study. *Genome Res*., **15**, 376–384.

Gascuel,O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol*., **14**, 685–695.

Giot,L. *et al*. (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.

Ito,T. *et al*. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Li,S. *et al*. (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**, 540–543.

Martin,D. *et al*. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol*., **5**, R101.

Rual,J.-F. *et al*. (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, **437**, 1173–1178.

Shannon,P. *et al*. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.

Stelzl,U. *et al*. (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.

Uetz,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.