BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 20 2011, pages 2919–2920
doi:10.1093/bioinformatics/btr485

Data and text mining

Advance Access publication August 19, 2011

Detecting biological network organization and functional gene orthologs

Jike Cui[1], Todd F. DeLuca[1], Jae-Yoon Jung[1] and Dennis P. Wall[1,2,*]

[1]The Center for Biomedical Informatics, Harvard Medical School and [2]Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA 02115, USA

Associate Editor: Jonathan Wren

ABSTRACT

SUMMARY: We developed a package TripletSearch to compute relationships within triplets of genes based on Roundup, an orthologous gene database containing >1500 genomes. These relationships, derived from the coevolution of genes, provide valuable information in the detection of biological network organization from the local to the system level, in the inference of protein functions and in the identification of functional orthologs. To run the computation, users need to provide the GI IDs of the genes of interest.

Availability: http://wall.hms.harvard.edu/sites/default/files/tripletSearch.tar.gz

Contact: dpwall@hms.harvard.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on June 24, 2011; revised on August 9, 2011; accepted on August 10, 2011

1 INTRODUCTION

A logic triplet of genes is a set of three genes with a logic relationship; for example, gene *c* is present in a genome if and only gene *a* and *b* are both present (Bowers *et al.*, 2004). We developed a method to effectively search for those triplets by incorporating phylogeny balancing into the phylogenetic profiles of triplets. We found logic triplets to be very powerful in the detection of biological network organization and identification of key genes and their protein functions, including adjacent and distant proteins in a pathway, and even interactions among different pathways at the system level (Cui *et al.*, 2011). Here, we provide the programs to search those logic triplets based on a large phylogenetic profile matrix containing 182 bacteria genomes and 87 441 orthologous gene clusters. The matrix was generated through Roundup (DeLuca *et al.*, 2006), one of the largest gene orthologs databases on the internet, containing 770 bacteria, 55 archaea and 76 eukaryotes.

Among possible logic relationships involving three genes, the four types in Figure 1 are most likely to be present in real cellular networks (Bowers *et al.*, 2004; Cui *et al.*, 2011). In addition, we also developed methods to detect *if* instead of *if and only if* (*iff*) relationships for the four types, because our study suggested the former might be very common (Cui *et al.*, 2011). Please see 'List of *iff* logic relationship types', 'Search for *iff* relationships' and 'Search for *if* relationships' in Supplementary Material for details.

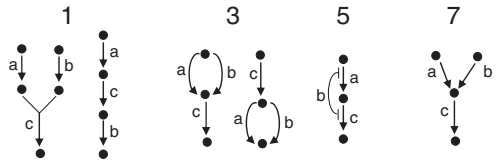

Fig. 1. The four basic logic relationships and their corresponding network organizations; black dots represent chemical compounds. Type 1: *c* is present in a genome iff *a* and *b* are both present. Type 3: *c* is present in a genome iff *a* is present or *b* is present. Type 5: *c* is present in a genome iff *a* is present and *b* is absent; it means protein *b* may inhibit enzyme *a* or *c*. Type 7: *c* is present in a genome iff one of either *a* or *b* is present. Terminology: iff means if and only if.

2 PROGRAM INTERFACE

The program package TripletSearch includes a file for the phylogenetic profile matrix of 182 representative bacteria genomes and 87 441 genes and 7 python and C++ programs. A python script, computeTriplets.py, provides a unified interface to users.

The script requires three inputs: the name of the phylogenetic matrix file, the name of the file to store triplets and the GI IDs of target genes.

The script also takes five optional inputs for users to fine-tune the number of triplets produced. (i) The minimum number of 0's or 1's a gene profile should contain. The default is 10% of the number of genomes in the matrix file; smaller numbers cause a gene profile to have too little entropy to be informative. (ii) The minimum number of 0 or 1 combinations in the joint distribution between the profiles of two genes. The default is 5% of the number of genomes in the matrix; smaller numbers not only lead to more triplets but also a higher false positive rate. (iii) The minimum ΔU value, a variable measuring the likelihood of triplets, ranging from 0 to 1 with increasing likelihood; default is 0.3. Please see 'Search for *iff* logic relationships' in Supplementary Material for details. (iv) The α value measuring the likelihood of *if* triplets for one tailed Z test; default is 0.99. And (v) the number of target genes a candidate triplet should contain. For example, if the status of a single gene in the whole cellular network is of interest, all the triplets containing that gene might be valuable. However, if the interaction of two genes is important, only triplets containing both genes may be wanted.

*To whom correspondence should be addressed.

Finally, a user might only focus on triplets comprised entirely of target genes.

3 APPLICATION

As our study suggested, logic triplets can be very helpful in the detection of biological network organization at different levels (Cui *et al.*, 2011). For example, a predicted logic triplet, *fabH is present iff either fabF or fabB is present,* matches exactly to the relationship of the three genes in *Escherichia coli*. In fatty acid biosynthesis, if *fabH* is responsible for the initiation of elongation, then the subsequent rounds of elongation can be carried out by either *fabF* or *fabB* (Keseler *et al.*, 2009) (See Supplementary Figure S1 for details).

In addition, logic triplets reveals the function and network location of putative genes. A triplet *tdcE is present iff both ybiW and pflD are present* indicates that enzymes coded by ybiW and pflD function closely with the one coded by *tdcE*. Indeed, although there is not yet any experimental data on those two genes, other computational studies have suggested such a link as well (Chen *et al.*, 2011; Keseler *et al.*, 2009).

Interestingly, logic triplets are also suggestive of translation regulation. A triplet *lldD is present iff one of either accA or accD is present* may seem odd at first. In *E.coli*, lldD catalyzes the production of pyruvate, which is then converted to actyl-CoA. accA and accD are the two subunits of acetyl-CoA carboxyltransferase. However, a recent study suggests that accD binds to the mRNAs of accA and accD to inhibit the translation and that such binding is mutually exclusive with the substrate binding for catalytic activity (Meades *et al.*, 2010). Therefore, the suggestion from the triplet that those two subunits cannot coexist in some bacteria is possible.

Finally, logic triplets can identify functional orthologs, the gene pairs with similar function but no sequence similarity. For example, an influential paper studied the formation of disulfide bond in bacterial proteins (Dutton *et al.*, 2008). While most bacteria contain DsbA and DsbB which function sequentially to form the bond, several major bacterial phyla do not have DsbB. The study finally identified VKOR as the functional ortholog of DsbB through experiments. With our method, a strong triplet *DsbA is present iff either DsbB or VKOR is present* would pinpoint the missed gene quickly.

In summary, logic triplets are useful in the detection of biological network organization, the inference of protein function and the identification of functional orthologs. Our package TripletSearch provides the phylogenetic profile matrix of large number of genes across a wide spectrum of bacteria genomes and a group of user-friendly programs to search for those triplets. Users can also create customized phylogenetic matrix from >1500 genomes available at our Roundup website (http://roundup.hms.harvard.edu/retrieve/).

Funding: We gratefully acknowledge support from the National Science Foundation under (grant number 0640809 to D.P.W.).

Conflict of Interest: none declared.

REFERENCES

Bowers,P.M. *et al.* (2004) Use of logic relationships to decipher protein network organization. *Science*, **306**, 2246–2249.

Chen,Y. *et al.* (2011) Genome-wide discovery of missing genes in biological pathways of prokaryotes. *BMC Bioinformatics*, **12** (Suppl. 1), S1.

Cui,J. *et al*. (2011) Detecting Biological Network Organization and Functional Gene Orthologs. *Bioinformatics*, PubMed PMID 21856738 [Epub ahead of print].

DeLuca,T.F. *et al.* (2006) Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, **22**, 2044–2046.

Dutton,R.J. *et al.* (2008) Bacterial species exhibit diversity in their mechanisms and capacity for protein disulfide bond formation. *Proc. Natl Acad. Sci. USA*, **105**, 11933–11938.

Keseler,I.M. *et al.* (2009) EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res.*, **37**, D464–D470.

Meades,G. Jr *et al.* (2010) A tale of two functions: enzymatic activity and translational repression by carboxyltransferase. *Nucleic Acids Res.*, **38**, 1217–1227.