BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 23 2011, pages 3325–3326
doi:10.1093/bioinformatics/btr584

Systems biology

Advance Access publication October 20, 2011

PRINCIPLE: a tool for associating genes with diseases via network propagation

Assaf Gottlieb[1,*], Oded Magger[1], Igor Berman[1], Eytan Ruppin[1,2] and Roded Sharan[1,*]

[1]The Balavatnik School of Computer Science, Faculty of Exact Sciences and [2]Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

Associate Editor: Jonathan Wren

ABSTRACT

Summary: PRINCIPLE is a Java application implemented as a Cytoscape plug-in, based on a previously published algorithm, PRINCE. Given a query disease, it prioritizes disease-related genes based on their closeness in a protein–protein interaction network to genes causing phenotypically similar disorders to the query disease.

Availability: Implemented in Java, PRINCIPLE runs over Cytoscape 2.7 or newer versions. Binaries, default input files and documentation are freely available at http://www.cs.tau.ac.il/~bnet/software/PrincePlugin/.

Contact: roded@tau.ac.il; assafgot@tau.ac.il

Received on August 16, 2011; revised on September 23, 2011; accepted on October 16, 2011

1 INTRODUCTION

Associating diseases with their causal genes is a fundamental challenge in medical research with applications to diagnosis and therapy. Recently, we introduced a novel method for prioritizing candidate disease-causing genes, named PRINCE (PRIoritizatioN and Complex Elucidation) (Vanunu *et al.*, 2010). PRINCE is motivated by the observation that genes causing similar diseases often lie close to one another in a protein–protein interaction (PPI) network (Oti and Brunner, 2007; Oti *et al.*, 2006). Given a query disease, PRINCE: (i) identifies a set of phenotypically similar diseases (van Driel *et al.*, 2006); (ii) retrieves the known causal genes of these diseases to form a ranked prior vector Y based on their similarity to the query and (iii) propagates the scores of the prior set of genes over a human PPI network to provide association scores for all genes. The final score assigned to each protein in the network combines the prior information with a network-based component. The latter ensures that the resulting scores are smooth over the network. Formally, the score $F(v)$ of a node v with a set of network neighbors $N(v)$ is:

$$F(v)=\alpha\left[\sum_{u\in N(v)} F(u)w(v,u)\right]+(1-\alpha)Y(v)$$

Where w is a normalized matrix representing the weighted PPI network and $Y(v)$ is the prior weight of the node. Here α is parameter weighting the relative importance of the prior-based versus the network-based components of the score.

PRINCE leverages on a comprehensive set of weighted PPIs compiled from multiple sources (Vanunu *et al.*, 2010), the disease–disease similarity measures computed by van Driel *et al.* (2006), and on the disease–gene associations presented in the Online Mendelian Inheritance in Man (OMIM) knowledgebase (Hamosh *et al.*, 2002).

Here we introduce PRINCIPLE (PRINCe ImPLEmentation)—a Cytoscape plug-in (Shannon *et al.*, 2003) implementation of the PRINCE algorithm. Given a query disease, it provides a list of top ranking genes associated with it and an additional visualization of the subnetworks formed by these top ranking genes and their direct interacting neighbors.

2 FUNCTIONALITY AND IMPLEMENTATION

The PRINCIPLE plug-in works in a client–server architecture, where a prior set of causal genes is propagated over the human PPI network, compiled from Breitkreutz *et al.* (2008); Ewing *et al.* (2007); Rual *et al.* (2005); Stelzl *et al.* (2005); Xenarios *et al.* (2002), residing on a designated server.

The PRINCIPLE plug-in includes three sections, represented in three tabs: (i) specifying the input files; (ii) specifying three tunable parameters that govern the algorithm scores and output size (see below); and (iii) specifying an optional output file listing the resulting network nodes. In the input files section, the query disease is selected from a sorted list of OMIM diseases (either by name or MIM code). A textual search for the query disease is also available. Three additional inputs are required: (i) OMIM phenotypic disease–disease similarity; (ii) a map file between MIM codes and disease names; and (iii) associations between diseases (MIM codes) and genes (Entrez IDs). While any user defined disease–disease similarities are applicable, the binaries page provides default choices for all. The default input files are described in the documentation page and include: (i) the phenotypic disease–disease similarity of Van-driel *et al.* (2006), which was also used in Vanunu *et al.* (2010); (ii) a disease names file corresponding to the similarity file entries (supplied by default with the plug-in); and (iii) a default set of disease–gene associations, extracted from GeneCards (Rebhan *et al.*, 1998), used also by Vanunu *et al.* (2010).

The PRINCIPLE plug-in provides three tunable parameters: (i) the weighting parameter $\alpha\in[0,1]$ (see Formula 1, with a default value of $\alpha=0.9$); (ii) $k\in(0,100]$, the number of top ranked genes to return (default 10); and (iii) $t\in(0,20]$, the number of iterations performed by the algorithm. The score $F(v)$ can be analytically

*To whom correspondence should be addressed.

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at TEL AVIV UNIVERSITY on December 18, 2011

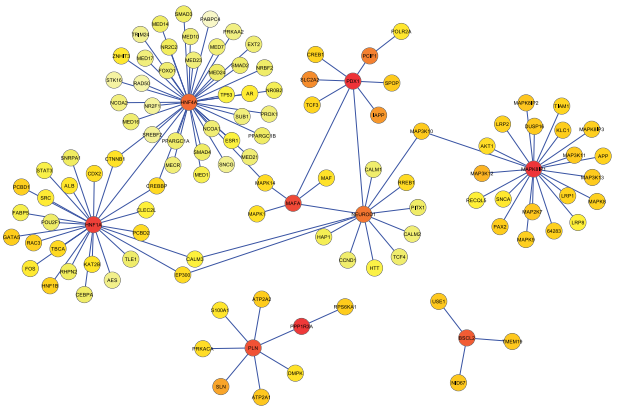

Fig. 1. An example of the PRINCIPLE output subnetwork for NIDDM, displaying an extract of the top 10 scoring genes and their immediate neighbors. Nodes are colored according to their association scores, with darker colors denoting higher scores.

Table 1. Top 10 associated genes and possible references to Diabetes

Rank	Gene	Supporting reference	Rank	Gene	Supporting reference
1	PDX1	OMIM	6	PLN	Bergha.A *et al.* (2006)
2	MAPK8IP1	OMIM	7	BSCL2	Chen *et al.* (2009)
3	PPP1R3A	OMIM	8	HNF4A	Moller *et al.* (1997)
4	HNF1A	Winckler *et al.* (2005)	9	NEUROD1	Liu *et al.* (2007)
5	MAFA	Kaneto *et al.* (2008)	10	PCIF1	Claiborn *et al.*

solved, but for efficiency we compute it using an iterative procedure (Zhou *et al.*, 2004). Typically, the algorithm shows fast convergence, achieving optimal results after 10 iterations (Vanunu *et al.*, 2010).

The results are displayed as the k top priority genes and their direct PPI neighbors, using a color scale signifying relative scores. An optional output file can be specified, listing the gene scores.

3 USAGE EXAMPLE

Figure 1 shows a typical output for querying Diabetes mellitus, non-insulin-dependent (NIDDM) (MIM 125853) with the default parameters ($\alpha = 0.9$, $k = 10$ and $t = 10$). The red circles are the top scoring proteins and their immediate PPI neighbors. These top 10 genes are listed in Table 1 along with references to articles studying their connection to Diabetes mellitus. Right clicking on a node enables retrieving additional information on the protein from multiple data sources.

ACKNOWLEDGEMENTS

We would like to thank Oron Vanunu and Tomer Shlomi for their help in the original PRINCE implementation. We would like to thank Nir Atias for helping with the plugin quality assurance.

Funding: Edmond J. Safra bioinformatics program (to A.G.); Bikura grant from the Israel Science Foundation (to E.R. and R.S.).

Conflict of Interest: none declared.

REFERENCES

Bergha,A,V.d. *et al.* (2006) Altered phosphorylation status of phospholamban and its contribution to the contractile dysfunction in mouse models of type II diabetes. *J. Mol. Cell. Cardiol.*, **40**, 925–926.

Breitkreutz,B.J. *et al.* (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.*, **36**, D637–D640.

Chen,W. *et al.* (2009) The human lipodystrophy gene product Berardinelli-Seip congenital lipodystrophy 2/seipin plays a key role in adipocyte differentiation. *Endocrinology*, **150**, 4552–4561.

Claiborn,K.C. *et al.* (2010) Pcif1 modulates Pdx1 protein stability and pancreatic beta cell function and survival in mice. *J. Clin. Invest.*, **120**, 3713–3721.

Ewing,R.M. *et al.* (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.*, **3**, 89.

Hamosh,A. *et al.* (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.

Kaneto,H. *et al.* (2008) PDX-1 and MafA play a crucial role in pancreatic beta-cell differentiation and maintenance of mature beta-cell function. *Endocr. J.*, **55**, 235–252.

Liu,L. *et al.* (2007) A novel mutation, Ser159Pro in the NeuroD1/BETA2 gene contributes to the development of diabetes in a Chinese potential MODY family. *Mol. Cell. Biochem.*, **303**, 115–120.

Moller,A.M. *et al.* (1997) Studies of the genetic variability of the coding region of the hepatocyte nuclear factor-4alpha in Caucasians with maturity onset NIDDM. *Diabetologia*, **40**, 980–983.

Oti,M. and Brunner,H.G. (2007) The modular nature of genetic diseases. *Clin. Genet.*, **71**, 1–11.

Oti,M. *et al.* (2006) Predicting disease genes using protein-protein interactions. *J. Med. Genet.*, **43**, 691–698.

Rebhan,M. *et al.* (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, **14**, 656–664.

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.

van Driel,M.A. *et al.* (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.*, **14**, 535–542.

Vanunu,O. *et al.* (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.*, **6**, e1000641.

Winckler,W. *et al.* (2005) Association of common variation in the HNF1alpha gene region with risk of type 2 diabetes. *Diabetes*, **54**, 2336–2342.

Xenarios,I. *et al.* (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.

Zhou,D. *et al.* (2004) Learning with local and global consistency. In *Advances in neural information processing systems 16: proceedings of the 2003 conference*, pp. 595–602.

Downloaded from http://bioinformatics.oxfordjournals.org/ at TEL AVIV UNIVERSITY on December 18, 2011