Bioinformatics Advance Access published January 28, 2011

PiNGO : a Cytoscape plugin to find candidate genes in biological networks

Michael Smoot [a], Keiichiro Ono [a], Trey Ideker [a] and Steven Maere [b,c*]

[a]Departments of Bioengineering and Medicine, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA. [b]Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium. [c]Department of Plant Biotechnology and Genetics, Ghent University, B-9052 Ghent, Belgium

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Summary: PiNGO is a tool to screen biological networks for candidate genes, i.e. genes predicted to be involved in a biological process of interest. The user can narrow the search to genes with particular known functions or exclude genes belonging to particular functional classes. PiNGO provides support for a wide range of organisms and Gene Ontology classification schemes, and it can easily be customized for other organisms and functional classifications. PiNGO is implemented as a plugin for Cytoscape, a popular network visualization platform.

Availability: PiNGO is distributed as an open-source Java package under the GNU General Public License (http://www.gnu.org/), and can be downloaded via the Cytoscape plugin manager. A detailed user guide and tutorial are available on the PiNGO website (http://www.psb.ugent.be/esb/PiNGO).

Contact: steven.maere@psb.vib-ugent.be

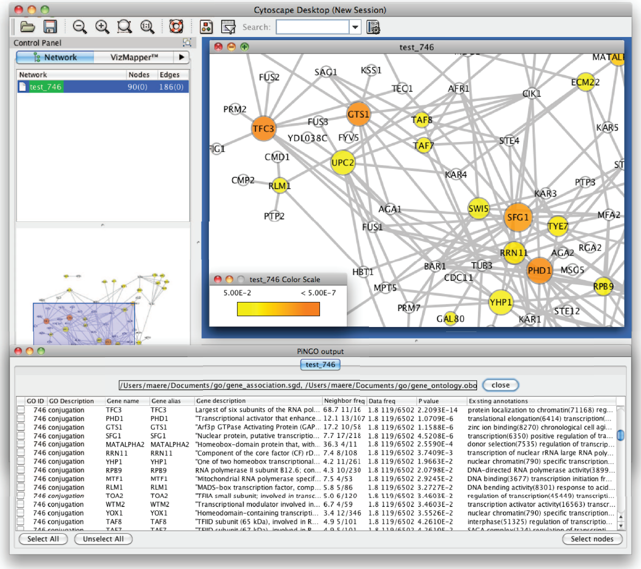
1 INTRODUCTION

A key problem for many molecular biologists is the identification of candidate genes to advance the study of a process or pathway of interest. A variety of strategies have been developed over the years to identify such candidate genes, mostly based on the guilt-by association principle. Two broad classes of methods can be distinguished (Sharan *et al.*, 2007): network-based methods (or direct methods) and module-based methods. In module-based methods, the data, for instance gene expression datasets or interaction networks, are clustered into modules which are functionally annotated using Gene Ontology (GO) (Ashburner *et al.*, 2000) or another functional categorization scheme. The functional annotation of a module is then transferred to its member genes. Some evidence however indicates that the module-based approach to predicting gene function may not be optimal. Wu *et al.* (2002) for example established that simply taking the top-10 correlated expression partners to predict the function of genes works better than traditional clustering methods. Similarly, Sharan *et al.* (2007) found indications that network-based methods outperform module-based methods, although a comprehensive comparison was not performed.

Among the network-based methods, simple first-neighbor based methods like the majority vote algorithm (Schwikowski *et al.*, 2000), where the function of a gene is predicted to be the most frequently occurring function among the gene's direct network neighbors, often yield surprisingly good results compared with more sophisticated methods involving propagation of functional information through the network (Nabieva *et al.*, 2005, Chua *et al.*, 2006, Murali *et al.*, 2006). Lossless propagation of functional annotations through the network, e.g. by considering the n-neighborhood of a gene with $n > 1$, generally gives rise to decreased performance, indicating that direct neighbors are the most relevant for predicting a gene's function (Sharan *et al.*, 2007, Nabieva *et al.*, 2005). A range of network-based function prediction methods have been developed that employ more sophisticated machine learning techniques (Sharan *et al.*, 2007 and references therein). These outperform naive methods but are computationally expensive, and because of this most of them have not been implemented as GUI-based tools (Sharan *et al.*, 2007). A couple of web-based tools, for instance AraNet (Lee *et al.*, 2009), ENDEAVOUR (Aerts *et al.*, 2006) and GeneMania (Warde-Farley *et al.*, 2010), allow users to prioritize candidate genes using probabilistic integrated networks. Although very useful, these tools are focused on one or a few organisms and, with the exception of GeneMania, do not allow users to upload their own datasets.

Despite the importance of candidate gene discovery for molecular biology research, only a limited number of flexible tools have been developed for this purpose. Here we present PiNGO, a user-friendly tool that answers questions of the type: 'Are there genes of class A in the network that are significantly connected to known class B genes but have no known role in process C ?'. The main use for PiNGO is to screen networks for novel genes that could be involved in a particular process. For instance, if one would be interested in discovering novel transcriptional regulators of conjugation in yeast, the above sentence would read: 'Are there transcriptional regulators in the network that are significantly connected to known conjugation genes but have no known role in conjugation ?' (see Figure 1). PiNGO is implemented as a plugin for Cytoscape, an open-source software platform to visualize, analyze and integrate molecular networks (Shannon *et al.*, 2003).

*to whom correspondence should be addressed



/Users/maere/Documents/go/gene_association.sgd, /Users/maere/Documents/go/gene_ontology.obo

	GO ID	GO Description	Gene name	Gene alias	Gene description	Neighbor freq	Data freq	P value	Existing annotations
□	746	conjugation	TFC3	TFC3	Largest of six subunits of the RNA pol...	68.7 11/16	1.8 119/6502	2.2093E-14	protein localization to chromatin(71168) reg...
□	746	conjugation	PHD1	PHD1	"Transcriptional activator that enhance...	12.1 13/107	1.8 119/6502	1.0709E-6	translational elongation(6414) transcription(...
□	746	conjugation	GTS1	GTS1	"Arf3p GTPase Activating Protein (GAP...	17.2 10/58	1.8 119/6502	1.1588E-6	zinc ion binding(8270) chronological cell agi...
□	746	conjugation	SFG1	SFG1	"Nuclear protein, putative transcriptio...	7.7 17/218	1.8 119/6502	4.5208E-6	transcription(6350) positive regulation of tra...
□	746	conjugation	MATALPHA2	MATALPHA2	"Homeobox-domain protein that, with...	36.3 4/11	1.8 119/6502	2.5590E-4	donor selection(7535) regulation of transcrip...
□	746	conjugation	RRN11	RRN11	"Component of the core factor (CF) rD...	7.4 8/108	1.8 119/6502	3.7409E-3	transcription of nuclear rRNA large RNA poly...
□	746	conjugation	YHP1	YHP1	"One of two homeobox transcriptional...	4.2 11/261	1.8 119/6502	1.9663E-2	nuclear chromatin(790) specific transcription...
□	746	conjugation	RPB9	RPB9	RNA polymerase II subunit B12.6; con...	4.3 10/230	1.8 119/6502	2.0798E-2	DNA-directed RNA polymerase activity(3899...
□	746	conjugation	MTF1	MTF1	"Mitochondrial RNA polymerase specif...	7.5 4/53	1.8 119/6502	2.9245E-2	DNA binding(3677) transcription initiation fr...
□	746	conjugation	RLM1	RLM1	"MADS-box transcription factor, comp...	5.8 5/86	1.8 119/6502	3.2727E-2	DNA bending activity(8301) response to acid...
□	746	conjugation	TOA2	TOA2	"TFIIA small subunit; involved in transc...	5.0 6/120	1.8 119/6502	3.4603E-2	regulation of transcription(45449) transcripti...
□	746	conjugation	WTM2	WTM2	"Transcriptional modulator involved in...	6.7 4/59	1.8 119/6502	3.4603E-2	transcription activator activity(16563) transcr...
□	746	conjugation	YOX1	YOX1	"Homeodomain-containing transcripti...	3.4 12/346	1.8 119/6502	3.5526E-2	nuclear chromatin(790) specific transcription...
□	746	conjugation	TAF8	TAF8	"TFIID subunit (65 kDa), involved in R...	4.9 5/101	1.8 119/6502	4.2610E-2	interphase(51325) regulation of transcriptio...

Fig. 1. A sample PiNGO analysis on an ENIGMA (Maere *et al.*, 2008) co-differential expression network learned from the budding yeast microarray compendium of Hughes *et al.* (2000). 'Conjugation' (GO:0000746) was specified as the only filter and target GO category, and 'transcription regulator activity' (GO:0030528) as the start GO category. Darker nodes are more significantly associated with (a subcategory of) conjugation. White nodes are known conjugation genes. The area of a colored node is proportional to the number of neighboring conjugation genes.

2 METHODS AND IMPLEMENTATION

PiNGO implements a simple network-based method to find genes associated with processes or pathways of interest. The network to be analyzed can be loaded through Cytoscape or from a text file. Input networks may be gene coexpression networks, protein or genetic interaction networks, or integrated networks. Edge weights are not taken into account. Given one or more target GO categories, PiNGO screens the network for genes whose direct neighbors are enriched for those functional categories at a chosen significance level. Subcategories of the target categories are also screened. PiNGO uses hypergeometric or binomial tests to calculate enrichment statistics, and Bonferroni or Benjamini-Hochberg FDR corrections to adjust the resulting p-values for multiple testing (Maere *et al.*, 2005).

The results are summarized in the PiNGO output window. The candidate genes for each target category are listed along with p-values and associated raw counts that give a good indication of the prominence of the target category in the candidate gene's neighborhood. When available, gene descriptions and known GO annotations of the candidate genes are reported to facilitate interpretation of the results and prioritization. In addition, an output network containing the candidate genes and their target GO annotated neighbors is extracted from the input network and displayed in Cytoscape. The output network reveals which genes contributed to the discovery of particular candidate genes. Output networks may contain denser areas or clusters of several candidate genes connected to the same neighbors. For Cytoscape input networks, all node and edge attributes and their visual mappings are preserved in the output networks. For networks imported from text, the adjusted p-values of the candidate genes and the numbers of target GO annotated neighbors are mapped to the node color and size in the output network, respectively (see Figure 1).

A unique feature of PiNGO is the capability to exclude genes with certain functional properties from the analysis, or to focus on genes with particular functions. Users are typically not interested in rediscovering genes that are already known to be involved in the target process or closely associated processes. These processes can be excluded by specifying them as 'filter categories'. People may also be interested primarily in discovering potential regulators of a given process, rather than effectors. In this case, categories like 'transcription factor activity' and 'protein kinase activity' can be specified as 'start categories', causing only genes annotated to these categories to be screened. Even when using filter categories, some not-so-novel genes may pop up in the candidate gene list, in the sense that their involvement in the chosen target process may already be known but not annotated in GO.

PiNGO offers unparalleled flexibility in the use of ontologies and annotations through its sister tool BiNGO (Maere *et al.*, 2005), which has been refactored for this purpose and has to be installed separately. BiNGO provides default GO and GOSlim ontologies and annotations for a wide range of organisms, from bacteria to plants and animals. PiNGO also supports the use of custom ontologies, both in OBO format and flat text format, and annotations. This allows the user to use non-GO classification schemes or to use PiNGO on non-model organisms. Multiple identifier types and synonyms are supported when using built-in annotations or annotation files from the GO Consortium. PiNGO also features GO evidence code filtering, automated remapping of annotations to GOSlim type sub-ontologies, and the possibility to specify custom reference gene sets against which functional enrichment is to be tested.

3 CONCLUSION

PiNGO implements a simple but efficient algorithm to find candidate genes in biological networks. PiNGO's main strengths are its user-friendliness and flexibility.

ACKNOWLEDGEMENT

MS, KO, and TI were supported by the US National Institutes of Health grant numbers RR031228 and GM070743. SM is a fellow of the Fund for Scientific Research-Flanders (FWO). We thank three anonymous reviewers and Thomas Van Parys for useful comments.

REFERENCES

Aerts, S. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat Biotechnol.*, **24**, 537-544.

Ashburner, M. *et al.* (2000) Gene Ontology: Tool for the unification of biology. *Nat Genet.*, **25**, 25-29.

Chua, H.N., Sung, W.K. and Wong, L. (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics* **22**, 1623-1630.

Hughes, T.R., *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**,109-126.

Murali, T.M., Wu, C.J. and Kasif, S. (2006) The art of gene function prediction. *Nat Biotechnol.*, **24**, 1474-1475.

Lee, I., *et al.* (2009) Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nat Biotechnol.*, **28**, 149-156.

Maere, S., Heymans, K., and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448-3449.

Maere, S., Van Dijck, P. and Kuiper M. (2008) Extracting expression modules from perturbational gene expression compendia. *BMC Syst Biol.* **2**:33.

Nabieva, E., *et al* (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, **21** Suppl 1, i302-i310.

Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of protein-protein interactions in yeast. *Nat Biotechnol.*, **8**, 1257-1261.

Sharan, R., Ulitsky, I. and Shamir, R. (2007) Network-based prediction of protein function. *Mol Syst Biol.*, **3**:88

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

Warde-Farley, D., *et al*. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38** Suppl, W214-W220.

Wu, L.F. *et al*. (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat Genet.*, **31**, 255-265.