BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 12 2014, pages 1789–1790
doi:10.1093/bioinformatics/btu092

Systems biology

Advance Access publication February 14, 2014

GUILDify: a web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms

Emre Guney†, Javier Garcia-Garcia and Baldo Oliva*

Departament de Ciencies Experimentals i de la Salut, Structural Bioinformatics Group (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona, 08003 Catalonia, Spain

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Determining genetic factors underlying various phenotypes is hindered by the involvement of multiple genes acting cooperatively. Over the past years, disease–gene prioritization has been central to identify genes implicated in human disorders. Special attention has been paid on using physical interactions between the proteins encoded by the genes to link them with diseases. Such methods exploit the guilt-by-association principle in the protein interaction network to uncover novel disease–gene associations. These methods rely on the proximity of a gene in the network to the genes associated with a phenotype and require a set of initial associations. Here, we present GUILDify, an easy-to-use web server for the phenotypic characterization of genes. GUILDify offers a prioritization approach based on the protein–protein interaction network where the initial phenotype–gene associations are retrieved via free text search on biological databases. GUILDify web server does not restrict the prioritization to any predefined phenotype, supports multiple species and accepts user-specified genes. It also prioritizes drugs based on the ranking of their targets, unleashing opportunities for repurposing drugs for novel therapies.

Availability and implementation: Available online at http://sbi.imim.es/GUILDify.php

Contact: baldo.oliva@upf.edu

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

Received on August 26, 2013; revised on December 13, 2013; accepted on February 8, 2014

1 INTRODUCTION

During the past decade, disease–gene prioritization has been central to research efforts in the field of human genetics. The promise of suggesting novel associations for genetic disorders with implications to therapeutical improvements has yielded a broad spectrum of computational tools (Kann, 2010; Tranchevent *et al.*, 2011). Special attention has been paid on using physical interactions between the products of these genes to associate them with diseases (Barabasi *et al.*, 2011; Ideker and Sharan, 2008).

Methods using protein–protein interactions (PPIs) exploit the 'guilt-by-association' principle over the network topology to uncover new disease–gene associations. The guilt-by-association principle suggests that the genes whose products (proteins) interact with the products of known disease genes are more likely to be disease genes (Aerts *et al.*, 2006; Lage *et al.*, 2007). Recently, we proposed three novel algorithms for genome-wide prioritization of disease genes using PPI networks and showed that a consensus method combining these algorithms improved the prioritization (Guney and Oliva, 2012) when using the disease–gene associations in Online Mendalian Inheritance in Man (OMIM) database (Hamosh *et al.*, 2005). Combined with genomics and proteomics data, the method has been successfully used to identify a gene driving metastasis to bone in breast cancer (Santana-Codina *et al.*, 2013).

Available network-based prioritization tools use either disease–gene annotations from OMIM database (Gottlieb *et al.*, 2011; Kohler *et al.*, 2008) or a set of genes provided by the user (Chen *et al.*, 2009; Kacprowski *et al.*, 2013; Warde-Farley *et al.*, 2010) as initial associations (seed genes). These tools typically output the prioritization for a set of candidate genes; genes lying under a given genomic interval, a set of user-provided genes or top ranking genes (several hundred at most). Furthermore, some of these tools are accessible only through Cytoscape (Saito *et al.*, 2012) as a plugin (Gottlieb *et al.*, 2011; Kacprowski *et al.*, 2013).

Publicly available biological data repositories can also be used for mining initial phenotype–gene associations required by network-based prioritization methods without restricting to any predefined phenotype or candidates. Considering the limited availability of convenient interfaces that bridge network-based prioritization algorithms, we present GUILDify, an interactome-based prioritization server for phenotype-based characterization. GUILDify retrieves initial phenotype–gene associations (seeds) via free text search on biological databases and ranks the proteins in the interaction network for their relevance to the phenotype. When the queried species is *Homo sapiens*, GUILDify provides a ranking of the drugs related to the phenotype based on the ranking of their targets.

2 NETWORK-BASED PRIORITIZATION USING INTEGRATED DATA SOURCES

The query made through GUILDify web interface is tokenized into keywords, and products of genes matching these keywords are searched in

*To whom correspondence should be addressed.
†Present address: Department of Physics, Center for Complex Network Research, Northeastern University, Boston, MA 02115, USA.

UniProt, OMIM and GO databases (see Supplementary Material on the data sources used). Matching proteins are displayed to the user. The proteins selected in this page are used as the initial phenotype–protein associations (seeds). The user also selects the prioritization method(s) to be used in the prioritization step. GUILDify maps the seeds (selected proteins) onto a genome-wide PPI network and runs the global topology-based prioritization algorithm selected by the user. The species-specific PPI network is generated using the interaction databases integrated in BIANA (Garcia-Garcia *et al*., 2010). GUILDify uses three algorithms (NetScore, NetZcore and NetShort) available to prioritize genes potentially involved in diseases using *a priori* disease–gene associations and PPIs. The user can modify the default parameters of the prioritization algorithms. Default values were tested on a large dataset of disease phenotypes using GUILD framework (Guney and Oliva, 2012). If more than one algorithm is selected, GUILDify uses a consensus that combines the scores of the selected algorithms (see Supplementary Material for details).

GUILDify outputs a likelihood score (GUILD score) associating the gene product with the phenotype provided by the user for each gene product in the PPI network. It outputs the rank and descriptive information of the gene products. The files containing this information and the seed proteins used in the prioritization can be downloaded from the web page displaying the results (results page).

The results page includes an interactive visualization panel where top-ranking proteins (at top 1 or 5%) and their interactions are displayed (Fig. 1). Nodes of the subnetwork in this panel can be selected, providing information on their description in BIANA integrated database. If the species is *H. sapiens*, GUILDify fetches the drugs (if any) targeting these top-ranking proteins from DrugBank (Knox *et al*., 2011) and displays them in the visualization panel. The score for each drug targeting a protein in the high-scoring subnetwork is calculated as the geometric distance to the protein(s) t targeted by the drug in the highest scoring subnetwork, using their individual scores as length. Further details on the materials and methods can be found in the Supplementary Material.

3 CONCLUSION

Phenotypic characterization of genes plays a crucial role in explaining the mechanisms behind biological processes. We have developed GUILDify, a free and easy-to-use web server for prioritization of genes using PPI networks. For a given phenotype, GUILDify uses descriptive fields in several proteomics and genomics databases in combination with network-based prioritization methods and provides an interactome-wide ranking. The ranking represents the relevance to the phenotype of interest and can be used to short-list the set of candidate genes that need to be further validated or to repurpose drugs (e.g. through common high-ranking targets).

ACKNOWLEDGEMENT

E.G. acknowledges the technical support from GRIB IT team, in particular A. Gonzalez and M. Sánchez Gómez.

Funding: 'Departament d'Educació i Universitats de la Generalitat de Catalunya i del Fons Social Europeu' through an FI fellowship granted to E.G.; by the grants from the Spanish Ministry of Science and Innovation (MICINN), FEDER BIO2011-22568, PSE-0100000-2007 and by EU grant EraSysbio+ (SHIPREC) Euroinvestigación (EUI2009-04018).

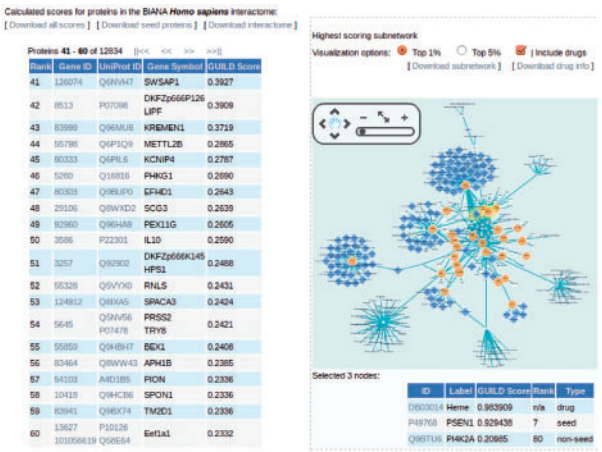

Fig. 1. Visualization panel in GUILDify prioritization results page using all the proteins retrieved with the keyword 'Alzheimer' as seeds and a combination of NetScore, NetZcore and NetShort algorithms

Conflict of interest: none declared.

REFERENCES

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

Barabasi,A.L. *et al*. (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet*., **12**, 56–68.

Chen,J. *et al*. (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*., **37**, W305–W311.

Garcia-Garcia,J. *et al*. (2010) Biana: a software framework for compiling biological interactions and analyzing networks. *BMC Bioinformatics*, **11**, 56.

Gottlieb,A. *et al*. (2011) PRINCIPLE: a tool for associating genes with diseases via network propagation. *Bioinformatics*, **27**, 3325–3326.

Guney,E. and Oliva,B. (2012) Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PLoS One*, **7**, e43557.

Hamosh,A. *et al*. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **33**, D514–D517.

Ideker,T. and Sharan,R. (2008) Protein networks in disease. *Genome Res*., **18**, 644–652.

Kacprowski,T. *et al*. (2013) NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*, **29**, 1471–1473.

Kann,M.G. (2010) Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief. Bioinform*., **11**, 96–110.

Knox,C. *et al*. (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*., **39**, D1035–D1041.

Kohler,S. *et al*. (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet*., **82**, 949–958.

Lage,K. *et al*. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotech*., **25**, 309–316.

Saito,R. *et al*. (2012) A travel guide to Cytoscape plugins. *Nat. Methods*, **9**, 1069–1076.

Santana-Codina,N. *et al*. (2013) A transcriptome-proteome integrated network identifies endoplasmic reticulum thiol oxidoreductase (ERp57) as a hub that mediates bone metastasis. *Mol. Cell. Proteomics*, **12**, 2111–2125.

Tranchevent,L.C. *et al*. (2011) A guide to web tools to prioritize candidate genes. *Brief. Bioinform*., **12**, 22–32.

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