BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 18 2014, pages 2689–2690
doi:10.1093/bioinformatics/btu344

Data and text mining

Advance Access publication May 20, 2014

MaxLink: network-based prioritization of genes tightly linked to a disease seed set

Dimitri Guala[1,2,*], Erik Sjölund[1,2] and Erik L. L. Sonnhammer[1,2,3]

[1]Stockholm Bioinformatics Centre, Science for Life Laboratory, SE-17121 Solna, [2]Department of Biochemistry and Biophysics, Stockholm University, Stockholm, SE-11321, Sweden and [3]Swedish eScience Research Center, SE-10450 Stockholm, Sweden

Associate Editor: Jonathan Wren

ABSTRACT

Summary: MaxLink, a guilt-by-association network search algorithm, has been made available as a web resource and a stand-alone version. Based on a user-supplied list of query genes, MaxLink identifies and ranks genes that are tightly linked to the query list. This functionality can be used to predict potential disease genes from an initial set of genes with known association to a disease. The original algorithm, used to identify and rank novel genes potentially involved in cancer, has been updated to use a more statistically sound method for selection of candidate genes and made applicable to other areas than cancer. The algorithm has also been made faster by re-implementation in C++, and the Web site uses FunCoup 3.0 as the underlying network.

Availability and implementation: MaxLink is freely available at http://maxlink.sbc.su.se both as a web service and a stand-alone application for download.

Contact: dimitri.guala@scilifelab.se

Supplementary information: Supplementary materials are available at *Bioinformatics* online.

Received on February 3, 2014; revised on May 1, 2014; accepted on May 12, 2014

1 INTRODUCTION

The vast amount of gene and protein interaction data emanating from high-throughput experiments have been skillfully compiled into networks of functional associations, such as FunCoup (Schmitt *et al*., 2014). These networks combine information on different types of protein–gene interactions from different species, data sources and data types to achieve a much higher quality and interaction coverage than pure protein–protein interaction networks. Owing to their high coverage, such networks have a high potential of yielding predictions about previously un-annotated gene function(s) and/or disease associations, through 'guilt-by-association' (GBA), i.e. assuming related functions of network neighbors (Lee *et al*., 2011).

MaxLink (Östlund et al., 2010) applies a GBA algorithm to search the FunCoup network to identify and rank genes enriched in connections to a seed set of known disease genes. In the original implementation, the algorithm was tested and validated on a set of cancer genes. The new version of MaxLink has been implemented as a web resource and has been further generalized to identify and rank potential disease genes of any disease or condition where a seed set of implicated genes can be provided.

*To whom correspondence should be addressed.

2 IMPLEMENTATION AND FEATURES

The original MaxLink Perl program has been re-implemented in C++, using the Boost library (http://www.boost.org/), for increased performance and a more general framework.

The new MaxLink implementation takes as input a list of genes, i.e. known to be involved in a particular disease or condition. The web service is run using default or user-specified values for the following parameters: 'Query network' (default: *Homo sapiens*)—one of the 11 currently available model organism networks in FunCoup 3.0 (Schmitt *et al*., 2014); 'Network confidence threshold' (default: 0.75)—a confidence threshold for the links in the underlying network, known as *pfc* in FunCoup (Alexeyenko *et al*., 2011);and the 'Hypergeometric Cutoff' (default 0.1)—a hypergeometric probability cutoff to ensure that identified genes are statistically enriched in connections to the query set. These parameters can be altered according to user guidelines found at the MaxLink Web site. After searching the FunCoup network, a list of candidate genes is returned. The list is sorted by the connectivity to the query list (MaxLink score). The resulting list can also be visualized using jSquid (Klammer *et al*., 2008) in the context of the searched FunCoup network (Fig. 1). As an example query, we used the nine breast cancer genes from the COSMIC cancer gene census database (Forbes *et al*., 2011) with the highest number of network links. Running MaxLink with the example query returns the set of genes most strongly associated to the query set in the context of the underlying Funcoup network.

The original implementation used a connectivity filter to only retain the genes with proportionally more connections to the query set than to other genes. Doing this avoided highly connected genes from being automatically counted as potential candidates solely based on their high number of interactions. In the new version, a statistically more sound approach has been implemented by calculating a hypergeometric probability score for each potential candidate based on its interactions with the query set and in total. Only candidates with a score below a cutoff are kept. This approach has been validated (Supplementary Material) on a set of rare disease genes from Orphanet (Aymé, 2003).

In the original implementation, an 'annotation filter' was applied to remove genes having annotations to cancer, as only

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

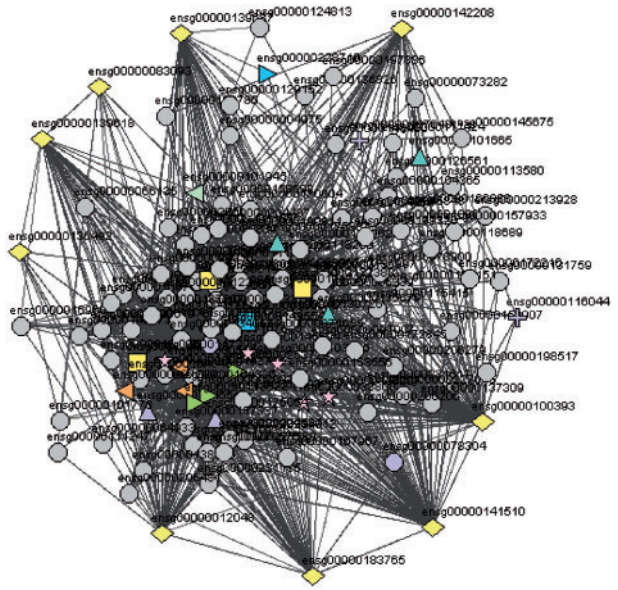

Fig. 1. jSquid graph of an example run of MaxLink, using default parameters and breast cancer genes from the COSMIC database (diamonds) as query. The genes identified by Maxlink in the FunCoup 3.0 network are significantly enriched ($P < 0.1$) in links to the query set, and thus predicted to be functionally important in breast cancer

novel potential cancer genes were of interest. This filter was removed in the new implementation to generalize the search to areas other than cancer. We here validate the new MaxLink on a set of genes associated with Alzheimer's disease (Supplementary Material) from the AlzGene database (Bertram *et al.*, 2007).

3 DISCUSSION

MaxLink, a network prioritization algorithm for genes, identified based on interactions with a seed set, has been improved by a statistically more sound specificity filtering. It is now released as a much faster C++ implementation, applicable to any set of seed genes, and available as an easy-to-use web resource. The tool can be used in general biological and applied biomedical research, e.g. prediction of potential disease genes, prioritization and/or analysis of results from GWAS/linkage studies.

Other tools exist for candidate gene prioritization, many of which are collected at the Gene Prioritization Portal (Tranchevent *et al.*, 2011). These tools use various strategies, e.g. text mining (Gonzalez *et al.*, 2008) or functional annotation analysis (Liekens *et al.*, 2011), and data sources, e.g. gene expression (van Dam *et al.*, 2012) and PPI networks (Köhler *et al.*, 2008). Most of the tools require a distinct region of the genome as input (Seelow *et al.*, 2008) to narrow down the search space. In contrast, MaxLink prioritizes genes from the whole genome, applying a unique GBA algorithm to FunCoup that integrates information on several types of protein and gene associations from many species and data sources, and is one of the most comprehensive networks available. Furthermore, many of the tools use keywords as input, as opposed to MaxLink, which only uses gene sets. Gene set-based tools have been shown to perform better than keyword-based ones (Börnigen *et al.*, 2012). Performance in general is, however, heavily dependent on the selection of evaluation datasets (Csermely *et al.*, 2013), but tools relying on more comprehensive data integration appear to be more powerful (Börnigen *et al.*, 2012). Performance benchmarking remains non-trivial owing to the diversity of the ways the tools are to be used, lack of solid experimental validation and knowledge contamination when statistical methods such as cross-validation are used (Moreau and Tranchevent, 2012).

A future extension of MaxLink could be to use other functional networks than FunCoup, to evaluate the algorithm on different data. This could also provide an even wider coverage of potential gene–protein functional interaction space. Furthermore, search methods such as shortest path (Peyer *et al.*, 2009) or PageRank (Brin and Page, 1998) could be implemented to exploit other topological properties of the network and to test the suitability of different types of search methods and queries.

ACKNOWLEDGEMENTS

The authors thank Gabriel Östlund and Thomas Schmitt for contributing with ideas to the development of this resource.

Funding: Premier Research Group, Merck AB.

Conflict of interest: none declared.

REFERENCES

Alexeyenko,A. *et al.* (2011) Comparative interactomics with Funcoup 2.0. *Nucleic Acids Res.*, **40**, 821–828.

Aymé,S. (2003) [Orphanet, an information site on rare diseases]. *Soins.*, 46–47.

Bertram,L. *et al.* (2007) Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.*, **39**, 17–23.

Börnigen,D. *et al.* (2012) An unbiased evaluation of gene prioritization tools. *Bioinformatics*, **28**, 3081–3088.

Brin,S. and Page,L. (1998) The anatomy of a large-scale hypertextual Web search engine. *Comput. Networks ISDN Syst.*, **30**, 107–117.

Csermely,P. *et al.* (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.*, **138**, 333–408.

Forbes,S.A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–D950.

Gonzalez,G. *et al.* (2008) GeneRanker: an online system for predicting gene-disease associations for translational research. *Summit Translat. Bioinforma.*, **2008**, 26–30.

Klammer,M. *et al.* (2008) jSquid: a Java applet for graphical on-line network exploration. *Bioinformatics*, **24**, 1467–1468.

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

Lee,I. *et al.* (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.*, **21**, 1109–1121.

Liekens,A.M.L. *et al.* (2011) BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol.*, **12**, R57.

Moreau,Y. and Tranchevent,L.-C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.*, **13**, 523–536.

Östlund,G. *et al.* (2010) Network-based Identification of novel cancer genes. *Mol. Cell. Proteomics*, **9**, 648–655.

Peyer,S. *et al.* (2009) A generalization of Dijkstra's shortest path algorithm with applications to VLSI routing. *J. Discrete Algorithms*, **7**, 377–390.

Schmitt,T. *et al.* (2014) FunCoup 3.0: database of genome-wide functional coupling networks. *Nucleic Acids Res.*, **42**, D380–388.

Seelow,D. *et al.* (2008) GeneDistiller–distilling candidate genes from linkage intervals. *PLoS One*, **3**, e3874.

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

van Dam,S. *et al.* (2012) GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics*, **13**, 535.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016