Bioinformatics, 31(12), 2015, 2052–2053
doi: 10.1093/bioinformatics/btv073
Advance Access Publication Date: 8 February 2015
Applications Note

OXFORD

Data and text mining

FUN-L: gene prioritization for RNAi screens

Jonathan G. Lees[1,*,†], Jean-Karim Hériché[2,†], Ian Morilla[3], José M. Fernández[4], Priit Adler[5], Martin Krallinger[4], Jaak Vilo[6], Alfonso Valencia[4], Jan Ellenberg[2], Juan A. Ranea[7] and Christine Orengo[1]

[1]Research Department of Structural & Molecular Biology, University College London, London, UK, [2]Cell Biology/Biophysics Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany, [3]Inflamex-Laboratoire Analyse Géométrie et Applications, Université Paris Nord-Sorbonne, France, [4]Structural Bioinformatics Group, Spanish National Cancer Research Centre (CNIO) and Spanish National Bioinformatics Institute (INB), Madrid, Spain, [5]Institute of Molecular and Cell Biology, and [6]Institute of Computer Science, University of Tartu, Tartu, Estonia and [7]Department of Molecular Biology and Biochemistry-CIBER de Enfermedades Raras, University of Malaga, Malaga, Spain

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Jonathan Wren

Received on November 2, 2014; revised on January 13, 2015; accepted on January 28, 2015

Abstract

Motivation: Most biological processes remain only partially characterized with many components still to be identified. Given that a whole genome can usually not be tested in a functional assay, identifying the genes most likely to be of interest is of critical importance to avoid wasting resources.
Results: Given a set of known functionally related genes and using a state-of-the-art approach to data integration and mining, our Functional Lists (FUN-L) method provides a ranked list of candidate genes for testing. Validation of predictions from FUN-L with independent RNAi screens confirms that FUN-L-produced lists are enriched in genes with the expected phenotypes. In this article, we describe a website front end to FUN-L.
Availability and implementation: The website is freely available to use at http://funl.org
Contact: ucbcjle@live.ucl.ac.uk

1 Introduction

An ever increasing volume of biological data provides insights into cellular biological processes. However, how best to leverage the data to generate experimentally testable hypotheses is an open issue. A case in point is RNAi screening where assay complexity, cost and available resources can limit capacity to conduct a large screen. When an exhaustive discovery of all genes involved in a biological process is not required, genome-wide screens are not necessary and a screen focused on candidate genes should be preferred. A key to the success of such screens resides in selecting genes with a high probability of being involved in the desired biological process. With the large amount and diversity of information now available, finding likely candidates has become non-trivial. Currently target selection for experiments is typically guided by experimentalists selecting a few candidates after manual database searches and extensively trawling the literature. This manual approach has the advantage of the experimentalists' ability to filter the quality of the data. However, this risks ignoring relevant experimental data provided by indirect evidence and from high-throughput experiments. Therefore, we developed Functional Lists(FUN-L) to help in selecting target genes for experiments. A practical use of FUN-L is in building focussed RNAi libraries.

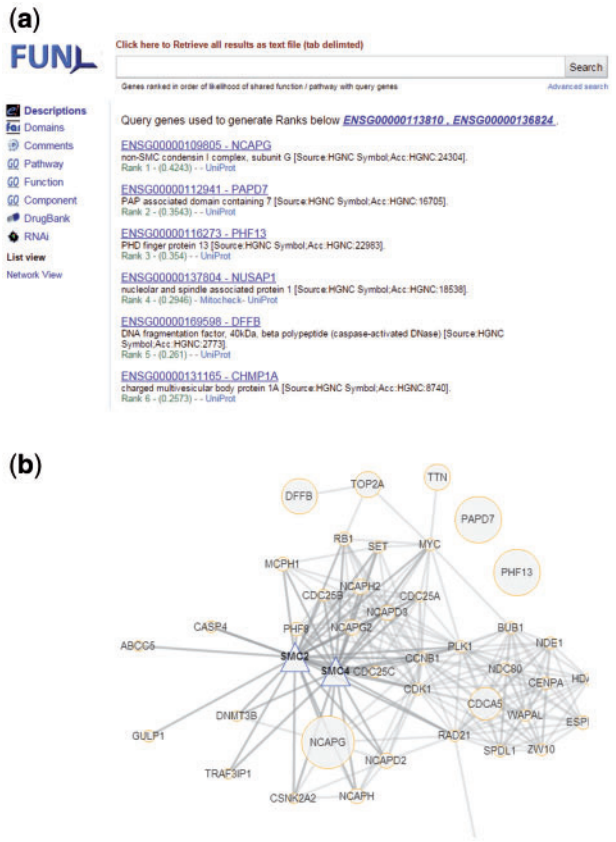

Fig. 1. (a) Example website search output shows a prioritized list of genes for testing and (b) the protein interactions with the query genes

2 Results

2.1 Algorithm and data sources

The data processing steps and algorithm used by FUN-L have been described elsewhere (Hériché *et al.*, 2014). In brief, FUN-L represents biological information from various sources as undirected weighted graphs from which measures of functional similarities between genes are derived using kernels on graph nodes. A distinguishing feature of FUN-L is that several kernel representations of the data were tested for information retrieval and only the best kernel/data combinations have been incorporated. As a result, FUN-L applies the commute-time kernel to the following data sources: GO similarities across biological processes, experimental protein interactions from other organisms mapped by orthology to the target organism, text mining from the iHOP (Hoffmann and Valencia, 2004) natural language processing protocol and the Random Forest kernel to experimental protein–protein interactions from the target organism. The FUN-L score for a gene is then simply the sum of its similarities to the query genes over the different data sources.

2.2 Website

The search interface is designed to be sparse and simple. Therefore, FUN-L is parameter-free, the only requirement is that the query is a set of genes representing the biological process of interest. As such, our method is akin to a search engine, where a set of input query terms returns a relevance ranking. In our resource, the search terms are gene identifiers rather than words.

The results of the search are also in a format similar to most search engines: a list of genes ranked by their score of relatedness to the genes used as query (Fig. 1a). The whole ranked list can be exported as a tab-delimited file for easy use in other applications such as siRNA design programs.

Further information such as biological process (Ashburner *et al.*, 2000), phenotypes (Neumann *et al.*, 2010; Schmidt *et al.*, 2013), gene essentiality (Chen *et al.*, 2012) druggability (Knox *et al.*, 2011) and cytoscape-web (Lopes *et al.*, 2010) protein interactions (Fig. 1b) can be displayed using links in the side bar.

3 Conclusion

The FUN-L website provides an easy-to-use interface to a validated biological process prediction algorithm. FUN-L provides a systematic, integrated, reproducible and fast tool for target prioritisation that is regularly updated. It compares favourably in performance with other applications, providing an independent parameter free method. Furthermore, it integrates extra unique information to aid in functional analysis of the proposed candidate list. We will develop these tools further in the near future to provide improved ability to interpret and filter the results. All data will be updated every 3 months to ensure results provided are up to date.

We anticipate FUN-L to be a useful tool in designing small libraries of interfering RNAs targeting a biological function of interest for cases where genome-scale screening is not affordable or possible. Importantly, the FUN-L method has been validated on independent data and has also been experimentally tested in human cells.

Funding

This work was supported by grants from the European Commission, Experimental Network for Functional Integration (Contract LSHG-CT-2005-518254), EU-FP7-Systems Microscopy NoE (Grant Agreement 258068), and EU-FP7-MitoSys (Grant Agreement 241548). I.M. and J.A.R were funded by SAF2012-33110 and CTS-486 (Spanish Ministry of Economy and Competitiveness, Andalusian Government and Fondos Europeos de Desarrollo Regional). The CIBERER is an initiative of the Carlos III Health Institute. J.G.L was part funded by BBSRC (Ref: BB/L002817/1).

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Chen,W. *et al.* (2012) OGEE: an online gene essentiality database. *Nucleic Acids Res.*, **40**, D901–D906.

Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nat. Genet.*, **36**, 664.

Hériché,J. *et al.* (2014) Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation. *Mol. Biol. Cell*, **25**, 2522–2536.

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

Lopes,C.T. *et al.* (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.

Neumann,B. *et al.* (2010) Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, **464**, 721–727.

Schmidt,E.E. *et al.* (2013) GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update. *Nucleic Acids Res.*, **41**, D1021–D106.