Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw508
Advance Access Publication Date: 2 August 2016
Application note

OXFORD

Genome analysis

NET-GE: a web-server for NETwork-based human gene enrichment

Samuele Bovo[1,†], Pietro Di Lena[2,†], Pier Luigi Martelli[1,*], Piero Fariselli[3] and Rita Casadio[1,4]

[1]Biocomputing Group, CIG, Interdepartmental Center «Luigi Galvani» for Integrated Studies of Bioinformatics, Biophysics and Biocomplexity, University of Bologna, Bologna, Italy, [2]DISI, University of Bologna, Bologna, Italy, [3]BCA, University of Padova, Padova, Italy and [4]Interdepartmental Center «Giorgio Prodi» for Cancer Research, University of Bologna, Bologna, Italy

*To whom correspondence should be addressed.
[†]These authors contributed equally to this work.
Associate Editor: John Hancock

Received on June 8, 2016; revised on July 18, 2016; accepted on July 20, 2016

Abstract

Motivation: Gene enrichment is a requisite for the interpretation of biological complexity related to specific molecular pathways and biological processes. Furthermore, when interpreting NGS data and human variations, including those related to pathologies, gene enrichment allows the inclusion of other genes that in the human interactome space may also play important key roles in the emergency of the phenotype. Here, we describe NET-GE, a web server for associating biological processes and pathways to sets of human proteins involved in the same phenotype
Results: NET-GE is based on protein–protein interaction networks, following the notion that for a set of proteins, the context of their specific interactions can better define their function and the processes they can be related to in the biological complexity of the cell. Our method is suited to extract statistically validated enriched terms from Gene Ontology, KEGG and REACTOME annotation databases. Furthermore, NET-GE is effective even when the number of input proteins is small.
Availability and Implementation: NET-GE web server is publicly available and accessible at http://net-ge.biocomp.unibo.it/enrich.
Contact: gigi@biocomp.unibo.it
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Big Data production in biomedicine is rapidly changing the way in which molecular knowledge is translated into health care (Bender, 2015). The spread and establishment of High Throughput Sequencing (HTS) technologies allows retrieving lists of interesting variations characterizing the investigated phenotype. In the context of functional genomics, each phenotype needs annotations for reconciling variations with known and putatively common biological processes and pathways, such as Gene Ontology (GO Consortium, 2015), KEGG (Kanehisa *et al.*, 2016), REACTOME (Fabregat *et al.*, 2016). At this level of biological complexity, a set of genes and their variations can acquire biological meaning and feature annotation only with an enrichment procedure (Laukens *et al.*, 2015). Enrichment helps in identifying within a set of genes some statistically significant and over-represented annotation features. Standard enrichment methods rely on the statistical over representation of the annotations that characterize the genes in the input set. Alternatively, network-based approaches extract graph properties from different interaction networks and pathways for modelling the complexity of the processes occurring in the cell and exploit this information for accomplishing the annotation enrichment in the context of protein functional interaction. Lists of web sites are available (Huang *et al.*, 2009; Laukens *et al.*, 2015; Mooney and Wilmot, 2015). Here, we introduce NET-GE, a web server that implements our method (Di Lena *et al.*,

2015), based on the extraction of subnetworks connecting proteins that share the same functional terms from a protein–protein interaction (PPI) network (Szklarczyk *et al.*, 2015). Differently from other methods also based on networks, our approach extracts modules that are function-specific by constructions and include all the seeds (proteins annotated with the same term) that in the PPI network are related to a specific functional annotation. One peculiarity of NET-GE is the possibility to enrich terms that are not present in the annotation of the starting protein set (and thus not detectable through a standard enrichment). When tested on the OMIM-derived benchmark sets, NET-GE is able to enrich sets of genes related to the same disease with biologically meaningful terms neglected by other methods (Di Lena *et al.*, 2015). The server, in addition, allows annotation based on KEGG and REACTOME pathways and a comparison between standard and network based enrichment.

2 NET-GE

NET-GE includes precomputed subsets of proteins associated to each functional terms of interest (Di Lena *et al.*, 2015). Subnetwork construction is based on the human interactome map downloaded from STRING (release 10.0, http://string-db.org/), or from a filtered version that retains only links with a score ≥ 0.9. Presently STRING includes 15,632 nodes (mapping 18 721 HGNC gene names, http://www.genenames.org/ and 89 085 UniProtKB identifiers) and 307 413 links (in the high quality STRING 0.9 version nodes and links are 9422 and 80 112, respectively).

In the present implementation of NET-GE, annotations are however available for all the 104 569 UniProtKB identifiers (release 2016_01), corresponding to 22 390 genes. The databases for annotating features are GENE ONTOLOGY (from UniProt-GOA human 145 resource, http://www.ebi.ac.uk/GOA); KEGG PATHWAY (release 77.0, http://www.genome.jp/kegg/pathway.html); REACTOME PATHWAY (release 53, http://www.reactome.org/). Redundancy among terms is not taken into consideration.

When generating the annotating subnetworks, for each annotation term we collect the seeds and evaluate the quality of the connecting nodes among seeds (for more details, see Supplementary Fig. 1S). After constraining seed distance, we determine the subset associated to a specific annotation term by retaining the minimal connecting subnetwork (Di Lena *et al.*, 2015). Considering STRING, NET-GE presently includes 20 391 annotation subsets (see http://net-ge.biocomp.unibo.it/enrich/statistics), 14 845 of which contain from two to 10 700 genes. The number of genes per subset is inversely proportional to the information content and the most informative terms correspond to small networks (Fig. 1).

The server implements both a standard and a network-based gene enrichment. Given a gene/protein list, each gene/protein is located in the different subsets of the annotation database. With a Fisher's exact test, the method estimates the overrepresentation significance of input genes/proteins in each precomputed subset for the corresponding annotation term. Standard enrichment includes only annotations of seed nodes; network-based enrichment includes seeds and their connecting nodes. For multiple testing correction, we use both the Bonferroni and the Benjamini-Hochberg (False Discovery Rate, FDR) procedures and evaluate a corrected p-value (Noble, 2009). Updating of the system, including human interactome and annotation databases is planned once a year, following the major releases.

2.1 Web server

NET-GE Web interface accepts UniProtKB Accession Numbers, Ensembl and HGNC gene names. The end user can select: (i) annotation modules based on STRING or STRING 0.9; (ii) the annotation (GO terms, KEGG, REACTOME); (iii) the multiple testing correction methods (Bonferroni or the Benjamini-Hochberg correction); (iv) the significance threshold.

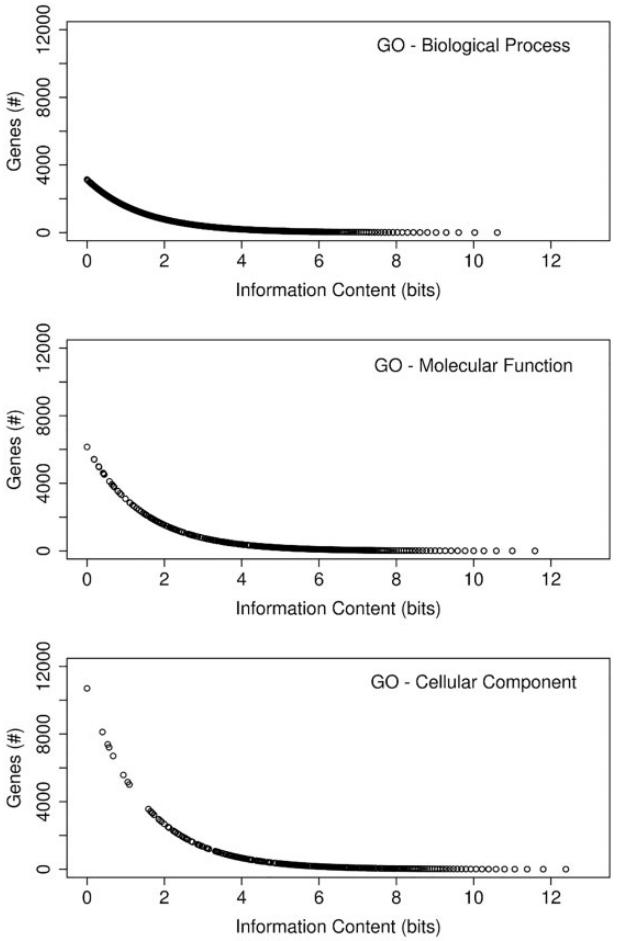

Fig.1. Dimension of subsets (number of genes) as a function of the information content for the Gene Ontology terms of the three main roots. The information content (in bits) is computed adopting standard methods (Shannon, 1948)

The output lists two enrichment tables: one for the standard and one for the network-based method (see the online tutorial for more details). Each table contains the annotation term identifier, linked to the corresponding database; the number and the list of input genes/proteins associated to the term; the *P*-value of the association; the description of the term and for the network based enrichment a visualization of the subnetwork. Enriched terms not included in the annotations of the input gene/protein are highlighted with a double star (see on line tutorial). It is also possible to access the complete set of annotations (for both the enrichment modes) of the submitted genes/proteins through the link provided at the bottom of the page. The front-end for the Web server follows the Model-View-Controller (MVC) paradigm, thanks to the web2py framework (http://www.web2py.com/), and it is optimized to work with all common web browsers. The analysis runs asynchronously: after submitting the query, the server displays a bookmarkable page reporting the status of the job. This page is periodically updated. A link to the results, accessible as soon as the job is completed, is given to the user. The final visualization of the results exploits the Graphviz library (http://www.graphviz.org/) and the JavaScript

library d3.js (http://d3js.org/). The user can also provide an e-mail address used to alert her/him as soon as results are ready. Running time depends on size of the input set (from two up to 200 genes) and ranges about 1–5 min.

Funding

RC thanks COST Action BM1405 (European Union RTD Framework Program) and FARB UNIBO 2012.

Conflict of Interest: none declared.

References

Di Lena,P. *et al*. (2015) NET-GE: a novel NETwork-based gene enrichment for detecting biological processes associated to Mendelian diseases. *BMC Genomics*, **16**, S6.

Fabregat,A. *et al*. (2016) The reactome pathway knowledgebase. *Nucleic Acids Res*., **44**, D481–D487.

Szklarczyk,D. *et al*. (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*., **44**, D447–D452.

Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res*., **43**, D1049–D1056.,

Huang,D.W. *et al*. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*., **37**, 1–13.

Kanehisa,M. *et al*. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*., **44**, D457–D462.

Laukens,K. *et al*. (2015) Bioinformatics approaches for the functional interpretation of protein lists: from ontology term enrichment to network analysis. *Proteomics*, **15**, 981–996.

Mooney,M.A. and Wilmot,B. (2015) Gene set analysis: A step-by-step guide. *Am. J. Med. Genet. B Neuropsychiatr. Genet*., **168**, 517–527.

Noble,W.S. (2009) How does multiple testing correction work? *Nat. Biotechnol*., **27**, 1135–1137.

Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Techn. J*., **27**, 379–423.