Protein Networks and Systems Biology
ELIXIR 2015, Bologna, Italy

Integration of PPI sources and prediction methods

NET-GE: a novel NETwork-based Gene Enrichment for detecting biological processes associated to Mendelian diseases

Samuele Bovo and Pietro Di Lena

Bologna Biocomputing Group, University of Bologna, Italy

Enrichment analysis and NET-GE, a summary...

Enrichment analysis is a tool for determining functional associations between biological processes/pathways and a set of genes/proteins related to the same phenotype. Enrichment analysis is useful for shedding light on the molecular mechanisms and functions at the basis of the analyzed phenotype and for enlarging the dataset of possibly related genes/proteins. Currently, several standard and network-based enrichment methods are available, but while standard methods rely only on the annotations that characterize the genes/proteins included in the input set, the network-based ones consider also the physical and functional relationships among different genes/proteins. Thus, considering a set of proteins in the context of their interaction network can help in better defining their functions. NET-GE is a novel NETwork-based Gene Enrichment tool for detecting the unifying biological processes and pathways at the basis of phenotypes. NET-GE implements a novel procedure based on the extraction from the STRING interactome of sub-networks connecting proteins that share the same annotations (Gene Ontology, KEGG and Reactome pathways). Enrichment analysis is performed by mapping the protein set to be analyzed on the sub-networks, and then by collecting the corresponding annotations. NET-GE is available at: http://net-ge.biocomp.unibo.it/enrich. Search options include Gene Ontology terms, KEGG or Reactome pathways and two different STRING networks. Results consist of significantly enriched annotations, also graphically depicted in the context of their relationships. For each annotation, proteins composing the module (networks) are listed. Network-based enriched annotations are emphasized in light-blue and those not-directly related to the input proteins are also aside reported.

NET-GE results: a brief description.

The page reporting the results can be divided in three areas:

1. A graph reporting the connection among the enriched terms. Boxes are color-coded considering the P-value of the enrichment. White boxes (connecting nodes) do not represent any enriched term but they are there in order to reconstruct/visualize the entire directed acyclic graph.
2. A main table reporting all the significantly the enriched terms, ranked by their P-Value.

These two tables consist of 7 different fields:

I) **Enrichment** - It reports if the term derives from the standard enrichment (S) or from the Network-based enrichment (N). If both methods enrich for a term, “NS” is reported; Network-based enriched terms are also colored in light blue.

II) **TERM** - It report the code of the term. The term is linked to the external related annotation database (QuickGO or KEGG or Reactome) containing detailed information about it;

III) **N1** - The number and the list of input proteins associated to the term. By clicking "[+] Show proteins" is possible to know the input proteins associated to the term;

IV) **N2** - The number and the list of proteins associated to the significant term. By clicking "Show proteins" is possible to list in another page the input proteins associated to the term; If the term is enriched for both methods, this set of proteins is referred to the network-based one.

V) **P-value** - A Bonferroni- or Benjamini-Hochberg- corrected P-Value

VI) **Description** - The name of the term;

VII) The link to the Term-specific network (only if available).

3. A second table listing only the enriched terms not included in the sets of annotations characterizing the input protein set. Columns are the same described above.

Please, note that....

Enrichment analysis can take minutes. Each job is queued and processed one at time. If you don't want to look the screen of your PC for minutes you can enter your e-mail address (last option in the main page). You will be e-mailed as soon as your results are available. To speed up the practicum, I provide the link to the results page.

Practicum 1 - OMIM #188890 TOBACCO ADDICTION, SUSCEPTIBILITY TO

Tobacco is one of the most abused substances in the world. Susceptibility to tobacco addiction has been linked to four genes SLC6A3, GABBR2, CHRNA4 and CYP2A6. Which are biological process are involved in this pathology?

Step 1

Retrieve from Uniprot/Swissprot the corresponding human Uniprot ACC (the reviewed one; option: filter by reviewed) . Describe briefly the function of these proteins. Example:

Gene name	Uniprot ACC (Human)	Function
SLC6A3	Q01959	Amine transporter

Step 2 – Enrichment analysis.

Enrichment analysis can be performed by using NET-GE (http://net-ge.biocomp.unibo.it/enrich.).

Fill the form as described. Use these parameters:

- Network of interaction: **String**
- Annotation database: **Gene Ontology – Biological Process**
- Correction Method: **Bonferroni**
- P-Value Threshold: **0.05**
- As suggested you can fill also the E-mail address field.

To speed up the practicum, results are available at:

http://net-ge.biocomp.unibo.it/enrich/read_out/2015_12_17_00_20_48/Bonferroni/0.05/nf/BP

Step 3 – Results – An overview

The figure about the DAG can give you an overview about the results...

Step 4a – Results - Standard method.

Considering the results based on the standard method (those having a “S” or a “N/S” in the first column). How many proteins of your list belong to those terms? Did you expect the involvement of these processes?

Step 4b – Results - Network-based method.

Considering the results based on the network-based method (those having a “N” in the first column), mow many proteins of your list belong to those terms? Did you expect the involvement of these processes? Look in PubMed (literature) to confirm some of them.

Can you appreciate the fact that the network-based method is able to add new information? Please, note that some processes are not included in the sets of annotations characterizing the input protein set... are you surprised by looking those results?

Practicum 2 - OMIM #188050 THROMBOPHILIA DUE TO THROMBIN DEFECT; THPH1

THPH1 is a disorder of impaired clot formation linked to four different genes: F13A1,F2, MTHFR and HABP2.Which biological processes are involved in this pathology? Can you appreciate the fact that the network-based method is able to add new information?

Results:

http://net-ge.biocomp.unibo.it/enrich/read_out/2015_12_17_00_59_48/Bonferroni/0.05/nf/BP