Gene Functional Annotation beyond Enrichment Analysis: moving from gene lists to functional metagroups and gene networks

Celia Fontanillo¹, Sara Aibar¹ and Javier De Las Rivas¹

¹Bioinformatics and Functional Genomics, Centro de Investigación del Cáncer (CiC-IMBCC, CSIC/USAL) Salamanca, Spain.

We present a computational method that filters the output of an Enrichment Analysis and finds significant coherent groups of genes and terms. These metagroups facilitate the interpretation of the biological functions and processes represented in a query gene list, inferring new associations among genes. The association of the genes to these “functional metagroups” is used to calculate a “functional score” for each gene. A comparison of these scores allows estimating a distance for each gene-pair. Using this method we build gene networks that reflect the proximity of each gene-pair based on a comprehensive analysis of their biological functional annotations.

Functional analysis of large sets of genes and proteins is becoming more and more necessary due to the increase of experimental biomolecular data at omic-scale. Enrichment analysis (EA) – like FEA or GSEA– is by far the most popular available methodology to derive functional implications of sets of cooperating genes. This type of analysis search in biological databases of gene attributes –such as Gene Ontology (GO) or KEGG pathways– and use statistical testing to find significant annotations assigned to specific genes. However this kind of methods do not address several issues related to the gene annotations, such us: redundancy of biological terms repeated in many databases (e.g. cell cycle GO:0007049, cell cycle KEGG hsa04110, etc), bias towards highly-frequent unspecific terms (e.g. GO:0050789 “regulation of biological process” includes more than 44% of all human genes annotated), or inadequate functional annotation of genes (e.g. NRAS is not annotated to GO:0043410 “positive regulation of MAPK cascade”).

To overcome these limitations we have developed a computational method that filters the output of an Enrichment Analysis and finds significant and coherent groups of genes and terms. These metagroups are able to summarize and facilitate the interpretation of the biological functions and processes represented in a initial query gene list, and also inferring new associations among genes that may be cooperating in a process which has not been annotated yet. The association of each gene to these “functional metagroups” is used, in a further step, to calculate a “functional score” for each gene. Based on the comparison of these scores we can estimate a “functional distance” for each gene-pair in the initial query gene list. Using this method we are able to build gene networks that reflect the proximity of each gene-pair based on a comprehensive analysis of their biological annotations.

The method developed has been tested with a small set of well-known interacting proteins from yeast and with a large collection of reference sets from three heterogeneous resources: mamalian multi-protein complexes (CORUM), yeast cellular pathways (SGD) and human diseases (OMIM). Moreover, we have also tested the algorithm with three sets of gene profiles published in specific experimental studies. Our algorithm, called “GeneTerm Linker” produces robust results even introducing different levels of noise in the genes used for the tests. The potential to provide gene networks based on biological functional annotations is also demonstrated in this communication.