Poster F-5

GOLIAS: Gene Ontology Library Analyzer for SAGE

Authors:
Gustavo GL Costa (*Universidade Estadual de Campinas- Hemocentro*)
Anderson Ferreira da Cunha (*Universidade Estadual de Campinas- Hemocentro*)
Tarcísio de Souza Peres (*Universidade Estadual de Campinas- Hemocentro*)
Tiago Ferraz Machado (*Universidade Estadual de Campinas- Hemocentro*)
Fernando Ferreira Costa (*Universidade Estadual de Campinas- Hemocentro*)

Short Abstract: Retrieving information from SAGE experiments demands integration among tools, databases and protocols. We developed GOLIAS, a program that uses the Gene Ontology DAG structure to provide this integration and to automate the SAGE analysis. GOLIAS reports quantitative data about the sequenced tags, statistics and graph charts.

Long Abstract:

Serial analysis of gene expression - SAGE - is a powerful method for generating global gene expression profiles for a cell or tissue type. Its main advantage is that it does not require prior information about the transcripts, thus, it can also be used for finding new genes. The SAGE technique, applied to a cell or tissue, returns a list of tags and counts, representing quantitative information about active genes. Each tag is 9 to 21bp long, depending on the SAGE protocol, and is supposed to uniquely identify a gene. Some tag-gene misassociations can occur, but they do not imply the average method efficacy loss. Informally, each count is an integer associated with the expression level of that gene. This is a raw measurement of gene expression. In a SAGE experiment, thousands of tags are sequenced. Typical numbers could be 20K unique tags sequenced and about 200K total tags, that is, the sum of all counts. In order to make conclusions regarding which processes are active in a cell or tissue, a greater stringency is desirable, that is, genes should be grouped in high-level categories. Gene Ontology (GO) [1] can help in this task. Gene Ontology is an international consortium designed to develop and maintain a hierarchy of English terms that describe many aspects of molecular biology. Gene ontology terms are not designed to be specific to any organism, but to be general. However, these terms do not describe any of the aspects related to metabolic pathways nor to gene expression. Its structure is a DAG (Directed Acyclic Graph), which is an ordinary directed tree, in which a node can have more than one parents, at different levels. As in a tree, directed cycles do not occur. GO's DAG structure allows assigning levels to each term and curators can annotate a biological entity at any level they want, depending on the amount of information about the entity they have. Gene Ontology will have a central role for SAGE experiment analysis. A number of software are available for gene expression experiments analysis using GO [David, Fatigo, Ease, GoMiner]. David is an on-line tool that provides as input a list of genes, one or more ontologies and one level. These input lists may be Unigene IDs, GeneBank accession numbers, etc. The GO analysis module assigns a GO term for each gene, from the ontology and level selected. Thus, GO terms from that ontology and level are listed with the number of genes inside each category. No information about the tags inside a category is output, neither could it be, since it does not form part of the input. A separate module of David, called Ease is available on-line and off-line. Ease was designed

to compare two libraries for differential expression with input from two lists: a background and a foreground. Background is the list of all genes in the experiment. Foreground is a list of differently expressed genes, previously selected by the user with his own criteria. Ease does not provide an interface to the foreground list selection; GoMiner, works in a similar fashion, although it receives a list of HUGO gene symbols as input and a integer number for each gene. This integer can be -1, 0 or 1, as the gene is up regulated, unchanged or down regulated. As output, the program produces a hierarchical list of Gene Ontology terms with statistical measurements to indicate which terms are changed. Although the tools described here are generic and can be used with many kinds of expression data, they do not consider aspects specific to SAGE experiments, such as total tags, unique tags and total genes assigned to each GO category. In addition, these programs do not take into account the statistical aspects of the comparison of two or more SAGE libraries. GOLIAS (Gene Ontology Library Analyser for SAGE) was developed in order to fill these gaps. The GOLIAS core was developed in Perl (www.perl.com) and its graphical user interface was developed in Delphi (www.borland.com/delphi/). Since the system was developed, it has been used in several SAGE projects performed in our laboratory (results in publication). As input, Golias takes a list of tags, tag counts and a Unigene Cluster ID as well as the level of Gene Ontology in which the user wants his data summarized. It generates a full report of GO terms along with quantitative tag information. The user can also see statistics about no matching, unique and total tags inside each GO term. Optionally, the user can see a GO pie chart view of one SAGE library, compare two libraries or see genes and descriptions inside each category.

[1] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. May;25(1):25-9.