A Web-Based Biological Data Mining Tool Based on Gen Ontology and Protein Interaction Pathways

Nashat Al-Jalad[1], Mahmoud Al-Saheb[2], and Yaqoub Ashhab[1]

[1]Biotechnology Research Center, Palestine Polytechnic University, P.O-Box 198, Hebron, Palestine
[2]Department of Information Technology, College of Administrative Sciences and Information Technology, Palestine Polytechnic University, P.O-Box 198, Hebron, Palestine

Introduction

The rapid advancement of bio-techniques as well as the introduction of various high-throughput laboratory technologies such as microarray and large scale proteomic experiments is generating an enormous amount of raw results that are considered meaningless data if not analyzed efficiently [1]. In the last two decades, biologists used to perform their routine bioinformatics analysis on a single biological entity such as a protein or a gene of interest. However, the way modern biology is explored rely heavily on system biology approach, where experiments are made to analyze and integrate the results of the whole set of genes, proteins, or the interconnections among them. Data mining-based statistical methods are playing a central role in facilitating knowledge extraction from large scale experiments. This work presents a powerful and easy to use web-based system that can perform analysis and representation of raw results which are usually obtained through high-throughput laboratory technologies.

Material and methods

Our system analyzes data and relationship between two sets of genes according to the Gene Ontology (GO) annotations and gene interaction pathways. The GO annotations were obtained from UniProt, which is a comprehensive resource for protein sequence and annotation data (www.uniprot.org). The analysis of gene interaction pathways were carried out based on Reactome database, which is an online curated resource for human pathway data that provides infrastructure for computation across the biological reaction networks (www.reactome.org) [2].

Two types of connections were used in our system. The first is a remote connection with untiprot database using MySQL connector-Net. The second is a connection to a locally Reactome database that was downloaded and converted from text mode to MS-Access database, while maintaining its structure in away to make it more accessible. Asp.net and VB.net language were used to build this system.

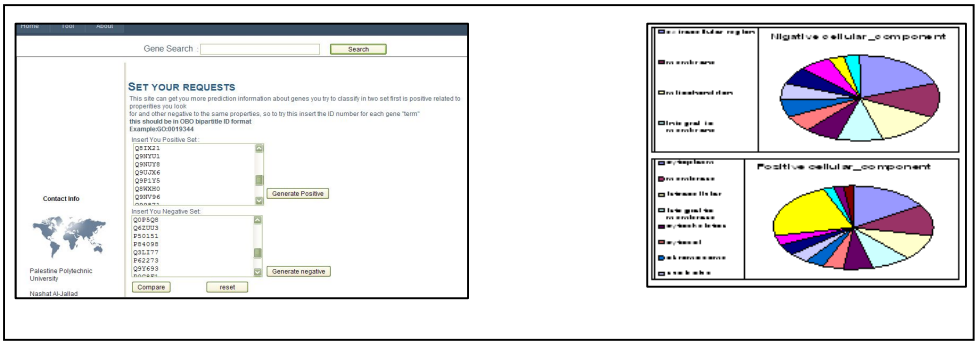
Our system is divides into three units. The first is a query unit that supports a powerful tool to browse a single gene profile. The second is called analyzing unit. It can discover common annotations and related functions between a set of genes as well as to perform annotations and functional comparison between two groups of unrelated data. The third is the reporting unit, which represents the analyses output graphically and textually in various useful formats.

The analyzing unit uses three terms of GO annotations; the Molecular function, the Cellular component, and the Biological process. Classification of text data is usually performed by establishing a number of sets in each group using scale-based sampling statistical methods [4]. The GO term rate per each group is used to find the normalized distribution of that term in a given set. The term rate per group is defined as follow: Let A and B be the genes from a group, V be the term of each gene, the equation is:

The results of this algorithm can be represented as two classes of proteins according to the biological treatment that was applied in the experiment; strongly associated and weakly associated proteins. To minimize background noise, the selected weakly associated GO terms from the first group is compared to all other GO terms in the second group. Those GO terms below the confidence threshold are automatically removed.

Results

Here we present the development of a user friendly we-based system for connecting, analyzing and presenting biological data that are usually generated as a result of large scale experiments of two treatment conditions. The user is asked to introduce the data of interest as protein accession numbers in two different windows (Figure 1a). Each accession number list should represents one type of treatment, e.g. genes unregulated by growth hormone versus genes down regulated by growth hormone. The user can select the output options such as Excel sheet or PDF graphical format (Figure 1b).



(A) (B)

Figure 1: A. System interface with the two windows to introduce proteins accession numbers. **B.** A typical graphical presentation of the GO terms analyses showing the positive and negative classes of the cellular component.

We tested our system using several real biological experiment outputs. The obtained results showed a good agreement with the results obtained from biological experiments. In most cases the correlation was over 80%. The system presented here is a strong tool that can predict relation between any two groups of genes without the need to perform any laboratory work.

References:

Myhre S, Tveit H, Mollestad T, Laegreid A. Additional gene ontology structure for improved biological reasoning. Bioinformatics. 2006. 15;22(16):2020-7.

Dotan-Cohen D, Kasif S, Melkman AA. Seeing the forest for the trees: using the Gene Ontology to restructure hierarchical clustering. Bioinformatics. 2009. 15;25(14):1789-95.

Matthews L et al,. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. 2009; 37(Database issue):D619-22.

Ramez Elmasri and Shamkant B. Navathe. Fundamentals of database system. Addison Wesley; 5 edition 2006.