Pubgen: Discovering and Visualising Gene-Gene Relations

Tor-Kristian Jenssen[1]
Tor-Kristian.Jenssen@idi.ntnu.no
Astrid Lagreid[2]
Astrid.Lagreid@medisin.ntnu.no
Jan Komorowski[1]
Jan.Komorowski@idi.ntnu.no
Eivind Hovig[3]
j.e.hovig@labmed.uio.no

[1] Department of Computer and Information Science,
Norwegian University of Science and Technology, Trondheim, Norway
[2] Department of Physiology and Biomedical Engineering,
Norwegian University of Science and Technology, Trondheim, Norway
[3] Department of Tumor Biology, Institute for Cancer Research,
The Norwegian Radium Hospital, Oslo, Norway

Keywords: gene-gene relations, visulalisation, Pubgen

1 Introduction

Many activities in functional genomics, particularly large scale gene-expression assays, give large quantities of data pertaining to a large number of different genes. Unsupervised methods for analysis, such as clustering algorithms, offers a number of ways to look at the data and summarize expression data, but still it remains difficult to interpret the results. It is not straightforward to compare two clustering analyses, and so far the only way to evaluate the output of a clustering algorithm is by manual inspection, heavily relying on detailed background knowledge [1]. As microarray technology evolves, the number of genes that can be printed on a single array rapidly grows to the 100k hypothesised to comprise the whole human genome, but the maximum number of genes about which a single researcher can have detailed knowledge is actually decreasing as the amount of detailed knowledge about each gene is increasing. Rapid and easy access to vital information about genes outside the biologists field of expertise will be decisive.

We have created a database of gene-gene relations by analysing Medline citation records. A pair of genes are related if they have been mentioned in the same article. Hence, the gene-gene links constitute a literature-derived network of genes. In addition to the gene-gene network, we have, by automatic means, obtained a list of literature references for a number of known human genes. By assuming that two genes are not mentioned in the same article by coincidence, but rather because there is some interesting relation between the two, the gene-gene network may provide important information aiding the analysis of expression data. We have constructed a web site offering a browsable presentation of the gene-gene network where the user may navigate the neighborhood of each gene*.

2 Methods and Results

We have used purely syntactic methods and searched titles and abstracts from Medline records to find articles describing human genes. By creating an index, or map, from each gene to a large number of articles, for all known human genes, and by reversing the map and combining the two maps, we obtain a map from gene to gene. The resulting map contains, for each gene, a list of links to all other genes with which the gene has been mentioned together.

The list of genes has been compiled by collecting and structuring data from public databases available through ftp or http. The main source of information has been the list of approved genes/symbols

*http://www.idi.ntnu.no/grupper/KS-grp/microarray/pubgen/genes.cgi from the HUGO Nomenclature Committee [2][†]. This list has been complemented by retrieving data from LocusLink[‡], The Genome Database (GDB)[§] , and GENATLAS[¶]. Adding information from these databases provided a number of gene entries, but, more importantly, provided a large number of literature aliases.

Altogether, including designations that have been withdrawn by HUGO but still used by other databases, we compiled a list containing 14961 genes. Based on a subset of Medline containing all publications from the years 1992 - 1999[‖], we found citations for 7561 (50%). Of these, 6978 had one or more co-citations (47%). The total number of articles was about 3 million, of which approximately 15% gave a hit for one or more genes. In the graphical display on our website the neighborhood of each gene is presented as a graph centered around the gene. Neighbors of the gene appear as nodes with edges weighted by the number of co-citations. The user can move around in the network by changing the gene in focus by clicking neighbors in the graph.

3 Discussion

Currently, the network contains a considerable amount of noise, i.e., incorrect gene to gene relationships. The signal-to-noise ratio may be enhanced by a number of means. One simple way is to apply threshold values to define valid relationships at the visualization stage. The most challenging task is to refine the criteria for accepting a gene as validly mentioned in an article. So far, we have used gene symbols as obtained from the gene databases and to some extent the gene names. A few particular genes and symbols require special considerations. One such gene is the Ii blood group gene, having the gene symbol II. The string "II", being used in a lot of other contexts than denoting the II gene, results in a large number of false hits for this gene[**].

However, our work clearly demonstrates that the concept is not only feasible, but has a number of downstream applications that may be exploited. For instance, such a network may be linked to human microarray experiments, where different representations may provide batch information on significant clusters being affected. Also, different visualisations of parts of the network may be selected based on specified criteria, such as type of protein. Eventually, a full three-dimensional representation may be built in a virtual reality setting for the user to explore various interactions. Further development strategies may include the use of natural language concepts for refinement of the type of interactions, i.e., inhibitive or potentiating etc., to further increase the usefulness of such a system.

Acknowledgements

We wish to thank Dyre Tjeldvoll for his help with the graphical software and the website. Also, we would like to thank The National Library of Medicine for giving us access to the Medline data.

References

[1] Iyer, V.R., *et al.*, The transcriptional program in the response of human fibroblasts to serum, *Science*, 283:83–87, 1999.

[2] White, J.A., *et al.*, Guidelines for human gene nomenclature, *Genomics*, 45(2):468–471, Oct 15 1997.

[†] `http://www.gene.ucl.ac.uk/nomenclature`

[‡] `http://www.ncbi.nlm.nih.gov/LocusLink/`

[§] `http://gdbwww.gdb.org/`

[¶] `http://www.citi2.fr/GENATLAS/`

[‖] The total number of articles in this subset was 3.1×10^6. We are continuing our indexing efforts and expect to complete all the 9 million records in Medline shortly.

[**] The II symbol also causes a large number of false hits for the phosphoribosyl pyrophosphate synthetase 2 (PRPS2) gene, for which the symbol is listed as an alias in the LocusLink database.