Vol. 00 no. 00 2005
Pages 1–2

The Gene Interaction Miner (GIM): A new tool for data mining contextual information for protein-protein interaction analysis

Aaron Ikin[1], Carlos Riveros[2], Pablo Moscato[2] and Alexandre Mendes[2,*]

[1]School of Electrical Engineering and Computer Science, The University of Newcastle, Callaghan, NSW, 2308, Australia.
[2]Research Centre in Bioinformatics, Biomarker Discovery and Information-Based Medicine (CIBM), School of Electrical Engineering and Computer Science, The University of Newcastle, Callaghan, NSW, 2308, Australia.

ABSTRACT

Motivation: This work was motivated by the need for an automated tool for discovery of genetic networks and the availability of extensive contextual protein-protein interaction information in the iHOP repository. At the moment, this information can not be explored to its full potential due to the lack of software tools to reliably collect, process and display that information in a way that life scientists can quickly analyze genes of interest and search for potential interaction networks. Commercial tools can perform a similar job, but results appear to be less informative than those obtained using contextual information.

Results: The Gene Interaction Miner (GIM) could successfully uncover complex network structures of protein-protein interactions for a test dataset composed of genes already related to Alzheimer's disease. That same set, when examined using two other analysis tools, namely STRING and Pathway Studio, resulted in incomplete protein-protein interaction networks, which indicate that the use of curated databases only gives a partial picture of the biological processes behind the disease.

Availability: The dataset used in this work and a running version of the software tool is available for download from the website http://www.cs.newcastle.edu.au/∼mendes/softwareGIM.html.

Contact: Alexandre.Mendes@newcastle.edu.au

1 INTRODUCTION

Most diseases with genetic backgrounds are triggered when one or more genes become faulty, creating a chain-effect in their regulatory networks or pathways. We expect those genetic networks and pathways to be reasonably reflected in scientific publications as studies frequently report such malfunctioning genes in batches. That would in principle validate the use of context-based data mining to find relations between genes, as reported in a recent review of tools for genomics analysis (Suderman and Hallett, 2007).

In terms of contextual data mining for protein-protein interaction research, the best repository at the moment is the iHOP (Information Hyperlinked over Proteins – http://www.ihop-net.org/), which has been cited over 1,500 times in the scientific literature since its creation in 2004 (Hoffmann and Valencia, 2004).

Among commercial tools that perform a similar task, we must cite *Pathway Studio* (Nikitin *et al*., 2003). Even though it performs in-depth functional analysis, among other features, the license for its use is too expensive for most institutions. Free online tools are also available and the most comprehensive is STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins* – see Jensen *et al*., 2009). There is a clear difference between GIM, Pathway Studio and STRING, though. GIM uses contextual information provided by iHOP, whereas Pathway Studio and STRING use their own curated databases. We will show that this has dramatic effects in the results.

2 APPROACH

In order to understand the type of context information stored in the iHOP repository, consider one of the contexts where BRCA2, a well-known breast cancer related gene, is cited:

> *"The breast cancer susceptibility protein BRCA2 controls the function of RAD51, a recombinase enzyme, in pathways for DNA repair by homologous recombination. [2002]"*

At first glance, such a context would immediately support a possible connection between the genes *BRCA2* and *RAD51*. However, if a large number of genes needs to be queried, or the number of contexts returned is too large, manual analysis quickly becomes inefficient. For instance, querying the gene BRCA2 alone in iHOP returns 398 contexts; and the task becomes too time-consuming and error prone.

The GIM automatizes this task by sequentially querying a user-defined list of genes of interest and then finding and counting the number of contexts in which two or more of them appear together. After that, it presents the results in a user-friendly manner, in which the user can, for instance, select specific gene-gene interactions; retrieve the publications that support them; and generate an output graph that allows a graphical visualization of the network of interactions between all genes queried.

*to whom correspondence should be addressed

3 EXAMPLE APPLICATION

A test dataset containing 32 genes related to Alzheimer's disease (see Brown *et al*. (2002)) was used to compare the three software tools. The network obtained by GIM was visualized using the yED software (http://www.yworks.com/) and is shown in Figure 1a. Each node represents a gene, and edges represent the publications that cited those genes within the same context. The edges' weights indicate numbers of publications.

The graph from STRING (see Figure 1b) is similar to the one obtained using GIM, but considerably more disconnected. There is no clear reason why STRING obtained such a disconnected graph, but it could be related to the criteria strictness for the inclusion of interactions in its curated database. However, a positive result is that the most connected genes in GIM are also connected in STRING.

Figure 1c shows the result for Pathway Studio. The graph contains a single edge and is considerably worse than the previous two. In this case, Pathway Studio has failed to provide any meaningful information. As before, we believe that Pathway Studio's strictness for the inclusion of relations in its curated database was responsible for the presence of a single edge. Because the test dataset contains genes already linked to Alzheimer's disease, one would expect the regulatory networks to be reflected somehow, but only GIM obtained such a result, followed by STRING, to some degree.

4 CONCLUSION

This paper presented the software tool GIM, which performs functional genomics analysis by using contextual information. It creates a gene-gene contextual citation graph that is then exported to a format that can be recognized by generic graph visualizers. GIM represents an alternative to other tools, such as Pathway Studio and STRING, which rely on curated databases to perform a similar task. Using a set of genes related to Alzheimer's disease as test data, GIM obtained a more structured interaction network. We point, however, that this might be due to the nature of the relations reported in the curated databases used by both STRING and Pathway Studio. If that is the case, GIM could indeed be a valuable tool for gene-gene interaction analysis, returning more meaningful information than other current packages.

ACKNOWLEDGEMENT

This work was supported by a *Competitive Independent Investigator* research grant provided by The University of Newcastle, Australia.

REFERENCES

Brown, V., Ossadtchi, A., Khan, A., Cherry, S., Leahy, R., Smith, D. (2002) High-Throughput Imaging of Brain Gene Expression, *Genome Research*, **12**, 244254.

Hoffmann, R., Valencia, A. (2004) A Gene Network for Navigating the Literature, *Nature Genetics* **36**, 664.

Jensen, L., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., von Mering, C. (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms, *Nucleic Acids Research*, **37**, D412-D416.

Nikitin, A., Egorov, S., Daraselia, N., Mazo, I. (2003) Pathway studio – the analysis and navigation of molecular networks, *Bioinformatics* **19**, 2155-2157.

Suderman, M., Hallett, M. (2007) Tools for visually exploring biological networks', *Bioinformatics* **23**, 26512659.

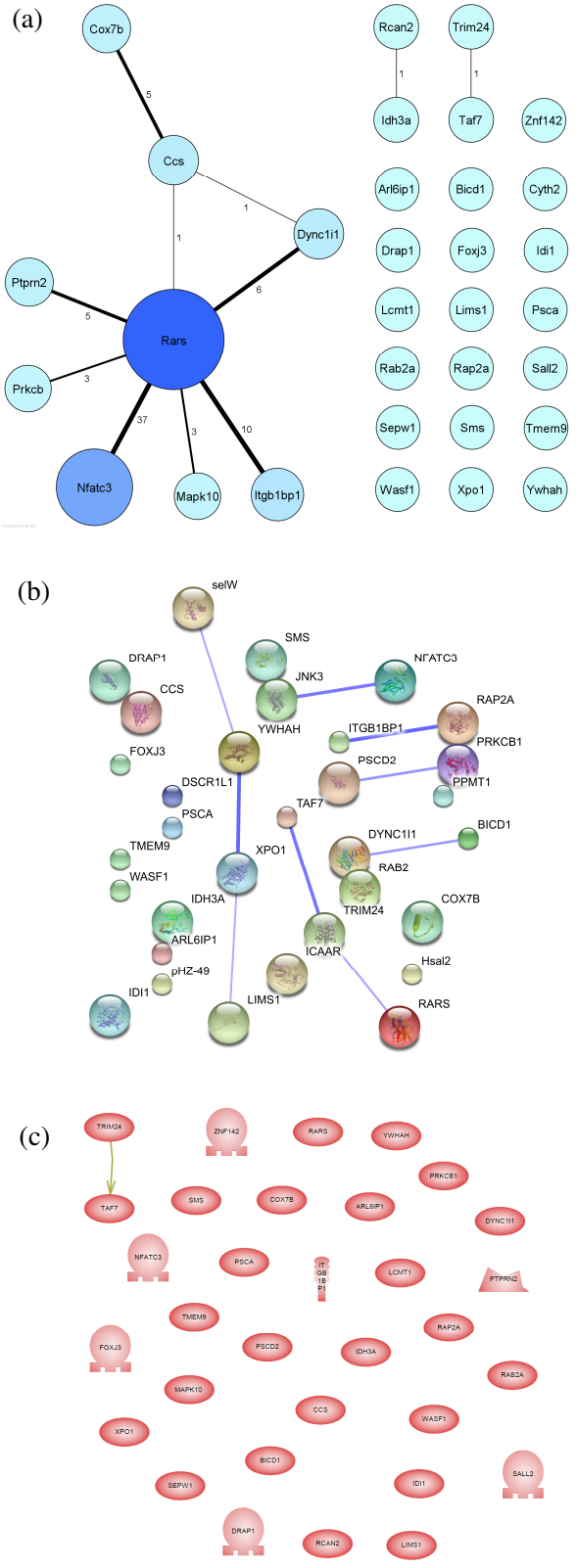

Fig. 1. (a) Network of context citations generated by GIM; (b) gene-gene interactions reported by STRING and (c) by Pathway Studio for the 32 genes differentially expressed in Alzheimer's disease reported in Brown *et al*. (2002).