



School of IT Technical Report



The University of Sydney

VISUALIZING THE GENE ONTOLOGY-ANNOTATED CLUSTERS OF CO-EXPRESSED GENES: A TWO-DESIGN STUDY TECHNICAL REPORT 622

DAVID CY FUNG AND SEOK-HEE HONG
SCHOOL OF INFORMATION TECHNOLOGIES
THE UNIVERSITY OF SYDNEY

KAI XU
NATIONAL ICT AUSTRALIA LIMITED

DAVID HART
AXOGENIC PROPRIETY LIMITED

APRIL 2008

Visualizing the Gene Ontology-Annotated Clusters of Co-expressed Genes: A Two-Design Study

David CY Fung¹, Seok-Hee Hong¹, Kai Xu², David Hart³

¹School of Information Technologies, The University of Sydney, Australia; ²National ICT Australia Limited;

³Axogenic Proprietary Limited

{dfun2647, seokhee.hong}@mail.usyd.edu.au, kai.xu@nicta.com.au, dhart@axogenic.com

Abstract-- In molecular biology, Gene Ontology (GO) has often been used for annotation and as a data mining dimension. A frequently performed step in microarray analytics is the clustering of co-expressed genes by their GO bioprocesses. Biological deductions are then made from the visual representation of the cluster pattern. Thus far, the question of how different representations of GO-annotated clusters affect biological interpretation and usability has not been investigated. In this paper, we evaluated two representations of GO-annotated clusters of co-expressed genes. Using a published cDNA microarray dataset, we tested the effect of each representation on biological reasoning. We also reported the results of the user evaluation conducted with bench biologists from different areas of expertise. Our study suggests that the bipartite graph may be more suitable for microarray analytics.

Keywords: bioinformatics, Gene Ontology, microarray, user evaluation, visualization design.

I. INTRODUCTION

In recent years, Gene Ontology (GO) has been used as a proxy for biological function and has increasingly been used for annotation and as a data mining dimension. Biologists often need to decipher the biological meaning of large microarray dataset by first determining the set of genes that co-expressed with each other and then by clustering the co-expressed genes according to their common GO bioprocess. Although GO visualizations are now widely used in microarray analysis software, the question on how different representations of GO-annotated clusters affect biological interpretation and usability has not been investigated. To answer this, we conducted a pilot study designed to compare two prototype representations, the *block matrix* and *bipartite graph*. The purpose was to find out which representation is more suitable for microarray analytics. We hope that this study will help bioinformaticians in designing visualizations that are usable to biologists. For the biologist, this study will help them to choose the visualization of relevance to their biological question.

II. RELATED WORK

Gene Ontology (GO) is a standardized set of controlled vocabulary for profiling the functional roles of genes and gene products (RNAs and proteins) in different species [1]. While many tools such as TreeMap [2] provide visualization on the parent-child relationship between GO terms, biologists are often more concerned with the visualization of gene-to-GO or gene cluster-to-GO relationships. This is because the cooperative regulation of bioprocesses is largely driven by the coordinated expression of genes [3].

HTP-GOMinerTM [4] and Exploratory Visual Analysis [5] provide visualizations of the gene-to-GO and cluster-to-GO relationships respectively while hiding the parent-child relationships between GO terms. The cluster map presented in HTP-GOMinerTM is a form of colour matrix designed to represent individual gene-to-GO relationship. The cluster pattern is formed by aggregates of coloured squares with no clear cluster boundaries. Exploratory Visual Analysis (EVA) is another form of colour matrix. The global cluster pattern is formed by a series of clearly bounded GO-annotated clusters arranged in a grid layout. Within each cluster is a matrix of nodes representing the subset of genes. Co-expressed genes are the nodes with the same colour hues. In EVA, the *n-ary* gene-to-GO relationship is being visualized as a 1:1 relationship leading to the same gene being drawn into multiple clusters. Therefore, unlike GOMinerTM, the visual semantics of EVA is GO-centric rather than gene-centric.

Bipartite graphs have often been used in metabolic pathway visualization in which the metabolites are represented by one set of nodes and the metabolic enzymes (a type of protein) that mediate the metabolic reactions are represented by another [6]. Representing GO-annotated protein interaction networks as a bipartite graph has also been used in which the GO terms are represented by one set of nodes and the proteins are represented by another [7]. Here we used bipartite graph to represent gene cluster-to-GO relationships.

In summary, the designing of visualizations depends on the biological relationships that need to be captured, the individual biologist's preference, and the objective of the biological question under study.

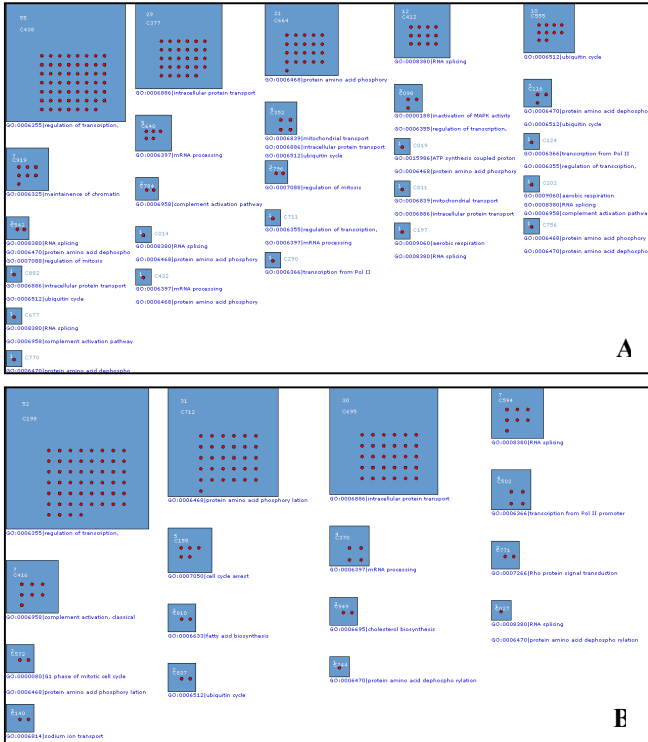


Figure 1. Visualization of co-expressed gene clusters as block matrix. (A) Normal liver; (B) Liver cancer

III. VISUALIZATION OF GO-ANNOTATED CLUSTERS

A. Block matrix

The block matrix is a form of matrix representation (Figure 1). Its visual semantics is gene-centric which emphasize on the functional partitioning of co-expressed genes using GO Bioprocess as the annotation. The purpose was to preserve the biologist’s gene-centric mental model.

In this representation, the n -ary gene cluster-to-GO relationship is being visualized as a 1: n relationship. It is visually simpler than any graph representation because it has no edge crossings or overlapping nodes. This design has each cluster of co-expressed genes being enclosed within a blue square. The cardinality of each cluster is drawn on the upper left-hand corner of the corresponding square. Each gene within a cluster is being represented as a circular node in red colour. This design requirement was recommended by two biologists in the preliminary user evaluation. While it lowers the information-to-ink ratio, the present design does fit the biologist’s mental model of a gene cluster. In their precept, a red node is an actively expressed gene and that a group of red nodes within a square means that the genes in the cluster are not only actively expressed but also positively co-expressed. This design is also supported by the ‘common field’ principle proposed by Chmielewski *et al.* [8]. It stated that the user tends to see a set of objects as a group if they are being drawn within an explicitly bounded, homogeneously coloured or textured region.

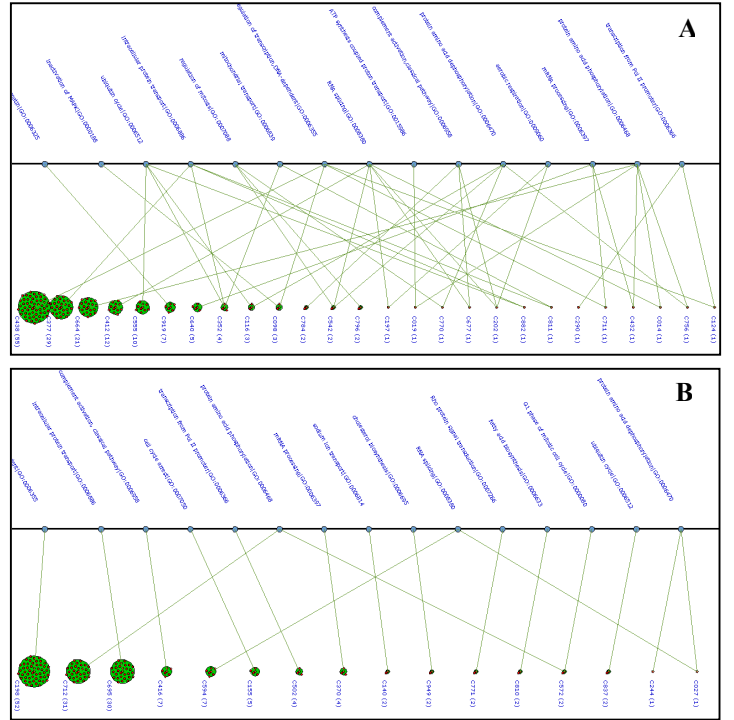


Figure 2. Visualization of co-expressed gene clusters as a two-level bipartite graph. (A) Normal liver; (B) Liver cancer

Unlike many grid layout algorithms [9], our design emphasizes on readability rather than optimum space usage. The design criteria are: (1) GO terms must be of a readable font size (10 points), (2) a maximum string size of 30 characters per GO term is allowed, (3) a visible spacing between gene clusters must be present, (4) each cluster must be clearly bounded, and (5) each cluster must be homogeneously coloured and in sharp contrast to the node colour.

The drawing algorithm involved four steps. The first step is to compute the grid formation (rows \times columns). This is done by rounding the square root of the number of gene clusters to the nearest integer. The second step is to compute the area of each cluster which is in direct proportion to the cardinality of the gene cluster. In the third step, the clusters are drawn in the descending order of their cardinalities across the grid from the left to the right. In the final step, the appropriate list of GO terms is drawn under each square.

B. Bipartite graph

The bipartite graph $G = (V_1 \cup V_2, E)$ contains two finite and disjoint sets of nodes V_1 and V_2 . E is a finite set of edges which is a subset of $V_1 \times V_2$ (Figure 2). Its visual semantics emphasize on the connectivity between V_1 and V_2 . The set of GO terms is being mapped to V_1 and the set of gene clusters is being mapped to V_2 . The gene-to-GO relationship is being mapped to E indirectly via the cluster-to-GO relationship. Each gene belongs to a single cluster and the n -ary gene cluster-to-GO relationship is being visualized as is.

The design criteria are the same as the block matrix. The bipartite graph is drawn in a two-level layout. The GO

nodes and the gene cluster nodes are assigned to the upper and the lower levels respectively. Each cluster of co-expressed genes is being enclosed within an elliptical node in green colour. The cardinality of each cluster is written in parentheses at the bottom of the corresponding circle in a vertical orientation. Each gene within a cluster is being represented as an elliptical node in red colour and GO nodes in blue colour.

The drawing algorithm involved five steps. The first step involved arranging the gene clusters in the descending order of their cardinalities from left to right in regular spacing. The second step is to compute the area of each cluster which is in direct proportion to the cardinality of the gene cluster. The third step is to pack the gene nodes into each cluster node using a phyllotactic layout [10] which has the advantages of spatial compactness and algorithmic simplicity. For the k -th gene node, it has the polar coordinates $(r, \Delta\theta)$.

And,

$$r = p\sqrt{k}$$

$$\Delta\theta = k \cdot 0.753 \text{ radians}$$

where p is the *packing factor*. The Cartesian coordinates of the k -th node can then be computed as:

$$x_k = x_0 + r \cos(\Delta\theta)$$

$$y_k = y_0 + r \sin(\Delta\theta)$$

where (x_0, y_0) is the Cartesian coordinates of the centre node and r is the radius of the plane at the k -th node. The fourth step is to minimize edge crossings by applying the barycenter algorithm once [11]. In the present case, the barycenter score $b(v)$ of every GO node v is defined as the arithmetic mean of the relative positions of its neighbouring gene clusters $N(v)$. Thus,

$$b(v) = \frac{1}{|N(v)|} \sum_{v \in N(v)} pos(v)$$

where $pos(v)$ is the relative position of a gene cluster to v . The GO nodes are then sorted in the ascending order of their barycenter scores and drawn from left to right in regular spacing. To complete the graph, cross-level edges are being added in the final step.

Both representations were implemented using the Java OpenGL library.

IV. CASE STUDY: LIVER CANCER

To evaluate the effect of the two representations on biological interpretation, the co-expression profiles of liver cancer and normal liver were used. These profiles, originally published by Chen *et al.* [12], were derived from 95 samples of primary liver cancer and 66 normal liver tissues. The data was later re-analyzed by Gamberoni *et al.* [13] who identified a set of co-expressed genes for each sample (disease or normal) based on their Pearson's correlation coefficients. The gene list provides pairwise co-expressed genes grouped according to their common GO bioprocesses. For experimentation, we extracted the GO bioprocess groups that belong to level 8 in the GO hierarchy. The gene list was first sorted by gene names.

This exposed the gene groups that share multiple GO bioprocesses. The remaining list was then sorted by GO bioprocesses. This exposed the gene groups that share a distinct GO bioprocess. We then assigned an alphanumeric label to each of the above groups with each mapping to a distinct cluster. In the following, the liver cancer dataset is referred to as the 'disease sample' whereas the normal liver dataset is referred to as the 'normal sample'.

Because identifying co-regulated bioprocesses is crucial to the understanding of gene regulation in cancer, a representation is useful to the biologist only if it can expose (1) the GO bioprocesses shared by the genes of the same cluster, (2) the GO bioprocesses shared by multiple gene clusters and (3) the distribution of GO bioprocesses in each sample. In the block matrix, one can easily identify item (1) by reading the listing of GO bioprocesses under each blue square. For example, in the normal sample, one could identify from the block matrix that cluster C542 contained a pair of co-expressed genes that are involved in protein amino acid dephosphorylation (GO:0006470), regulation of mitosis (GO:0007088), and RNA splicing (GO:0008380) (Appendix; Figure S1A). With the bipartite graph, the same has to be achieved by tracing edges which crossed at multiple points (Appendix; Figure S1B). However, there is no distinct advantage of block matrix over bipartite graph when it came to reading the disease sample. For example, it could be seen in the bipartite graph that cluster C572 in the disease sample contained two co-expressed genes that are involved in protein amino acid phosphorylation (GO:0006468) and G1 phase of mitotic cell cycle (GO:0000080) (Appendix; Figure S2B). However, the same were equally identifiable in the block matrix (Appendix; Figure S2A). Therefore, neither representation could be considered as a better alternative if very few clusters, in this case 12.5% of the total number of clusters, are related to more than one GO bioprocess.

The absence of edges in the block matrix facilitated pairwise comparison for functional relatedness between clusters. For example, it could be recognized that regulation of transcription (GO:0006355) is the bioprocess commonly shared between clusters C098 and C124 in the block matrix without the need to trace edges as with the bipartite graph (Appendix Figure S3). It is also noteworthy that the associated bioprocesses of GO:0006355 are different in each cluster. In cluster C098, inactivation of MAPK activity (GO:0000188) is the associated bioprocess whereas in cluster C124, transcription from PolIII promoter (GO:0006366) is the associated bioprocess. This is suggesting that the two clusters play different roles in transcriptional regulation but are functional complements of each other [14].

When attempting to identify the GO bioprocesses shared by more than two gene clusters in each sample set, the bipartite graph was more usable. That is because the representation of GO bioprocesses is non-redundant and the GO nodes are drawn at a different level from the gene clusters. To achieve the same with the block matrix would require one to tediously compare all possible pairs of clusters within each sample set. Thus the bipartite graph was better for identifying items (2) and (3).

Bipartite graph also allows one to identify differentially regulated bioprocesses between samples. This can be achieved by comparing the node degrees of the same GO bioprocess in different samples. For example, the GO node in the bipartite graph corresponding to ubiquitin cycle (GO:0006512) has a degree of four in the normal sample and has a degree of one in the disease sample. This immediately indicates that the bioprocess is switched off in liver cancer. Ubiquitin cycle is a protein degradation process known to be involved in the cell cycle arrest, DNA repair, and angiogenesis [15].

In summary, the block matrix seemed to be more suitable for examining the bioprocesses of an individual cluster and pairwise inter-cluster comparison for functional relatedness. The bipartite graph seemed to be more suitable for comparing between sample sets thereby exposing their biological differences. Since this is fundamental to hypothesis formulation, the bipartite graph would be better suited to microarray analytics than the block matrix.

V. USER EVALUATION

A. Participants

Our choice of participants emphasized on their quality as domain experts in biology. Fourteen participants were recruited from four medical research institutes and two university biology departments. They were experts in various fields of biology, e.g. biochemistry, cardiology, immunology, oncology, pharmacology, and virology. All of them were practicing bench biologists without formal qualifications in computer science.

B. Experimental setup

This experiment was setup to examine three independent variables and three between-subject dependent variables. The three independent variables are: microarray datasets (normal liver and liver cancer [13]), representations (block matrix and bipartite graph), and task types (competency and conceptual tasks). The three dependent variables are: task completion time, accuracy, and user confidence score. Accuracy is defined as the percentage of the total number of tasks being correctly answered. Both representations were presented as static graphics without any interactivity to ensure that the participant's performance was attributed only to the graphical form observed.

C. Procedure

Each session started with the evaluator explaining to the participant the design of the representation, the nature of each task and how to fill out the questionnaire. The participant was given a trial session to familiarize oneself with the procedure using a synthetic dataset. In the proper session, each participant performed tasks A to H twice, once on the normal sample and once on the disease sample. The person then performed tasks I and J once because they required both samples to be presented. For each task, the participant was timed, observations were gathered, and the

answers to the questionnaire collected. At the end of each session, the participant was asked to provide a subjective rating on a five-point scale (0 to 4) to indicate one's confidence in performing each task. The higher the score, the higher is the participant's confidence. The participant was also free to express one's opinion about the evaluated representation in writing.

In the following, we present the analytical task, the use case scenario of each task, and the summarized result of the evaluation. The group of participants in the block matrix evaluation is referred to as the 'block matrix group' and those in the bipartite graph evaluation are referred to as the 'bipartite graph group'.

D. Analytical tasks

Participants were given ten tasks. The first five (tasks A-E) were competency tasks designed to test the readability of each representation. The last five (tasks F-J) were conceptual tasks designed to test the effect of each representation on analytical reasoning. Based on our understanding of microarray analysis, we designed tasks that most microarray users will perform. The evaluation tasks and the use case scenario for each are listed in the following:

Task A. Find the gene cluster that is linked to the largest number of GO IDs.

Use case scenario. A biologist may want to identify co-expressed genes that are involved in multiple bioprocesses. A cluster linked to multiple bioprocesses is an indication that its member genes could be control points for coupling/decoupling various bioprocesses.

Task B. Find the gene cluster that is linked to the smallest number of GO IDs from each sample set.

Use case scenario. A biologist may want to identify co-expressed genes that are functioning in a particular biological pathway. A cluster linked to only one or two bioprocesses is an indication that its member genes are functionally specific.

Task C. Find the GO ID(s) that has the largest number of co-expressed genes from each sample set.

Use case scenario. A biologist may want to identify the bioprocesses that are the most active in the normal or the disease sample. A bioprocess with a larger number of co-expressed genes than others often indicates that it is relatively active.

Task D. Find the GO ID(s) that has the smallest number of co-expressed genes from each sample set.

Use case scenario. A biologist may want to identify the bioprocesses that are the least active in the normal or the disease sample. The rationale is the opposite of task C.

Task E. Find the GO IDs that are active only in the NORMAL or DISEASE sample set.

Use case scenario. A biologist may want to identify the bioprocesses that are specific to a particular phenotype. This task is fundamental not only to biomedicine but also to agriculture where biologists want to compare the biological difference between plant or livestock species.

Task F. Deduce which bioprocess is the most highly regulated.

Use case scenario. A biologist may want to identify bioprocesses that are highly regulated relative to others. Biologists interpret co-expression as synchronized activity among a group of genes and therefore must be co-regulated. The number of co-expressed genes linked to a bioprocess is an indicator of how highly regulated it may be.

Task G. Deduce which bioprocess is likely to be the least regulated.

Use case scenario. A biologist may want to identify which bioprocess(es) is the least regulated relative to others. The rationale is the opposite of task F.

Task H. Deduce which bioprocesses are likely to be co-regulated with the ubiquitin cycle (GO:0006512) and has the largest number of co-expressed genes.

Use case scenario. A biologist may want to use a particular bioprocess as a focus for investigating its connection with the other bioprocesses.

Task I. Deduce which human tissue the diagrams could most likely represent.

Use case scenario. This task was designed to test the usefulness of GO bioprocesses in biological deduction.

Task J. Deduce which disease could the DISEASE TISSUE most likely represent.

Use case scenario. This task was designed to test the usefulness of GO bioprocesses in biological deduction.

E. Evaluation results

The evaluation results were summarized in Figure 3. With the competency tasks, the median accuracy (Figure 3B) and the median user confidence score (Figure 3D) of the bipartite graph group was comparable with its block matrix counterpart. A comparison between the worst cases showed that participants in the bipartite graph group gave a 50% higher accuracy than those in the block matrix group (Figure 3B). Also, it took the block matrix group 40% longer to complete the competency tasks than the bipartite graph group (Figure 3A). Thus the advantage of bipartite graph over block matrix lies in faster task completion and, for some participants, reading accuracy. Some participants in the block matrix group said that the redundant representation of the GO bioprocesses made the normal sample confusing to read (Table 1) and could be the cause behind their prolonged task completion time.

Four participants in the bipartite graph group were finger tracing the edges to confirm any perceived relationships between a gene cluster and its neighbouring GO nodes or vice versa, but the same behaviour was absent with the block matrix group. The presence of edges seemed to give the participants a visual mean to confirm the presence of any perceived relationships. However, some participants found that the edge crossings in the normal sample set were interfering with graph reading (Table 1).

With the conceptual tasks, the median accuracy in the bipartite graph group was twice of the block matrix group (Figure 3C). However, the bipartite graph group took 36% longer to complete the conceptual tasks than the block matrix group (Figure 3A). That was because some participants in the bipartite graph group seemed to realize

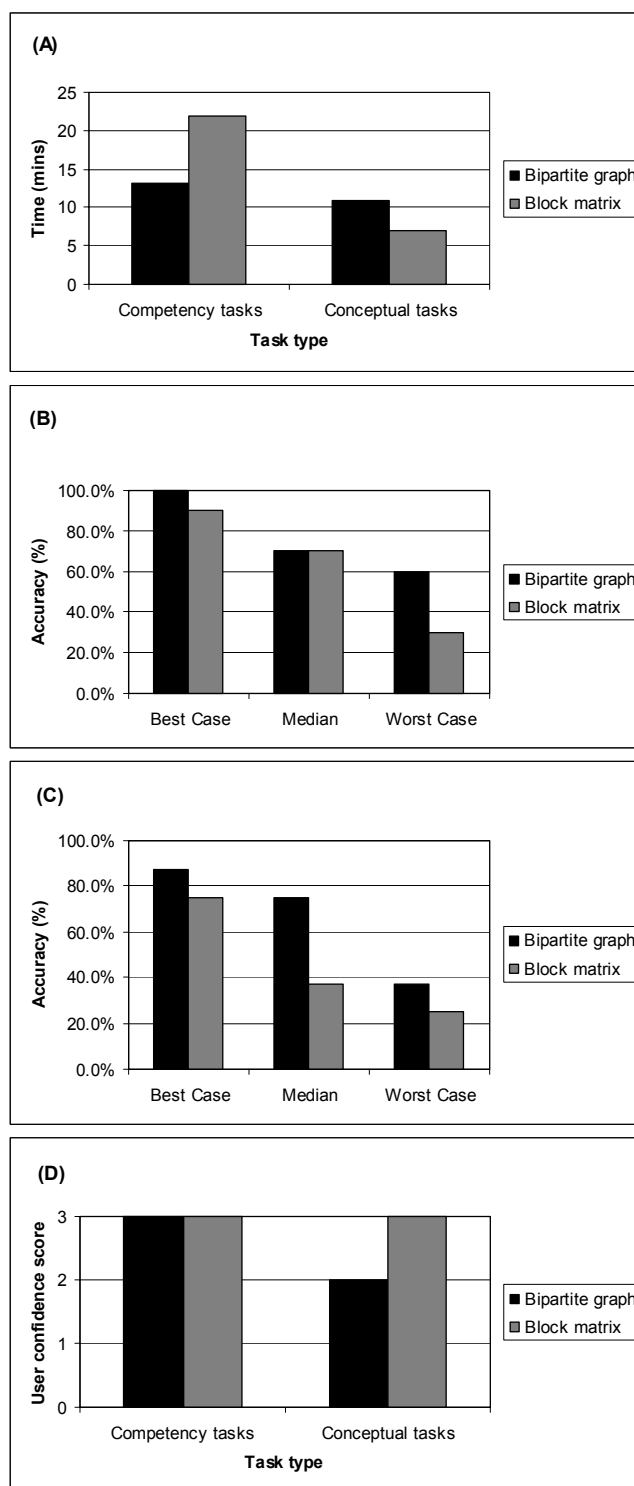


Figure 3. Summary of evaluation results. (A) Median task completion time per task type; (B) Overall median accuracy of competency tasks; (C) Overall median accuracy of conceptual tasks; (D) Median confidence score per task type.

the relatedness between some of the competency and conceptual tasks and were repeatedly re-examining not only the representation but also their answers made to the competency tasks. The median user confidence score showed that the bipartite graph group had a less positive user experience than the block matrix group (Figure 3D). This is suggesting a case of *perception/performance*

mismatch. In other words, participants in the bipartite graph group did not realize that they were giving a higher number of correct responses than its block matrix counterpart. Therefore, the advantage of the bipartite graph over block matrix lies in enhancing the biologist's analytical accuracy but in its current form is perceptually less usable than the block matrix. The participants' post-task comments reflected this (Table 1).

Of all the conceptual tasks, tasks I and J were the most challenging. Both tasks challenged the participants to make deductions based on their expertise in biology. Only one participant was able to answer both tasks correctly. Another two participants answered either task I or J correctly. All three participants were from the bipartite graph group. In terms of confidence scores, both groups gave a score of one indicating that they found tasks I and J difficult to perform. This may suggest that the GO bioprocesses presented are not informative enough for the participants to draw an accurate conclusion to either task. Furthermore, the answers given to tasks I and J seemed to be deduced from a selected few rather than from the entire set of GO bioprocesses. It was possible that the participants were exhibiting *cognitive bias* towards the bioprocesses that they are most familiar with. In the best case, one participant based his deduction for task I on four GO bioprocesses and another participant based his deduction for task J on three. If the participants had prior knowledge on the tissue type and the disease type of the microarray dataset as would be in the real-world scenario, this would suggest that a holistic understanding of liver cancer can only be achieved if the dataset has been cross examined by biologists from different areas of expertise.

VI. FUTURE WORK

With the user evaluation indicating that the bipartite graph is perceptually less readable but can enhance analytical accuracy more than the block matrix; several options can be taken to improve the usability of the former. Feasible options are listed as the follows:

1. To make the bipartite graph more informative, interactivity will be added to the gene nodes such as drop down menu on brushing to provide hyperlinks to public databases such as ENTREZ.
2. To further reduce the negative impact of edge crossings on readability, edge highlighting on pointer brushing can be added to the bipartite graph.
3. For comparing the cluster patterns of two sample sets (e.g. normal versus disease), a tripartite graph with the GO bioprocesses as the intermediate layer could be an option. This should enhance user performance on task E.
4. Another user evaluation can be performed by comparing a published visualization method with bipartite graphs in various designs. This should provide more information on their usability in microarray analytics.

VII. CONCLUSION

The strength of block matrix lies in its visual simplicity and its gene centric semantics. However, this apparent advantage over the bipartite graph did not translate into real performance enhancement in either the task completion time or in analytical accuracy. The underlying reason could be its redundant representation of the GO bioprocesses. Readability is further compromised when the redundant GO bioprocesses are scattered throughout the visualization.

The strength of the bipartite graph lies in its graphical semantics which emphasize on connectivity between two sets of nodes. It is a faithful depiction of the gene cluster-to-GO relationship. The better performance of the bipartite graph group in reading and analytical accuracy may imply that the graphical view is more relevant to biologists than the gene-centric view. However, the usability of bipartite graph decreases with the increase in edge crossings. Therefore edge crossing minimization is necessary for enhancing graph readability.

In conclusion, it seemed that the bipartite graph representation of GO-annotated gene clusters is more suitable to microarray analytics.

ACKNOWLEDGEMENT

We would like to thank Dr Rohan Williams for recruiting participants in the University of New South Wales and the Prince of Wales Hospital. We also thank all participants from the Ramaciotti Centre for Gene Function Analysis, Victor Chang Cardiac Research Institute, Viral Research Unit in the Prince of Wales Hospital, Westmead Millennium Institute, Drug Health Services (Research) in the Royal Prince Alfred Hospital, and the School of Biochemistry in the University of Sydney for taking part in this project. This project was supported by the Australian Research Council Linkage grant no. LP0455334.

REFERENCES

- [1] Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nuclei Acids Research* (database issue) 2006; **34**: D322-326.
- [2] E.H. Baehrecke, N. Dang, K. Babaria, B. Shneiderman. Visualization and analysis of microarray and gene ontology data with treemaps. *BMC Bioinformatics* 2004; **5**: 84-96.
- [3] Y. Qi, H. Ge. Modularity and dynamics of cellular networks. *PLoS Computational Biology* 2006; **2**: 1502-1510.
- [4] B.R. Zeeberg, H. Qin, S. Narasimhan, *et al.* High Throughput GOMiner, an 'industrial strength' integrative Gene Ontology tool for interpretation of multiple-microarray experiments, with applications to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics* 2005; **6**: 168-188.
- [5] D.M. Reif, M. Israel, J.H. Moore. Exploratory visual analysis of statistical results from microarray experiments comparing high and low grade glioma. *Cancer Informatics* 2007; **2**: 19-24.

- [6] R. Bourqui, L. Cottret, V. Lacroix, *et al.* Metabolic network visualization eliminating node redundancy and preserving metabolic pathways. *BMC Systems Biology* 2007; **1**: 29.
- [7] E. Demir. PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics* 2002 **18**: 996-1003.
- [8] T.L. Chmielewski, D.F. Dansereau, J.L. Moreland. Using Common Region in Node-Link Displays: The Role of Field Dependence/Independence. *Journal of Experimental Education* 1998; **66**: 197-207.
- [9] M. Bruls, K. Huizing, J.J. van Wijk. Squarified Treemaps. *Proceedings of the joint Eurographic and IEEE TVCG Symposium on Visualization* 2000, Eurographics Association; 33-42.
- [10] S. Carpendale, A. Agarawala. PhylloTrees: Harnessing nature's phyllotactic patterns for tree layout. *IEEE Symposium on Information Visualization* (Texas, USA) 2004, IEEE Computer Society Press: Chicago; 215.3.
- [11] K. Sugiyami, S. Tagawa, M. Toda. Methods for visual understanding of hierarchical system structures. *IEEE Transactions on Systems, Man, and Cybernetics* 1981; **SMC-11(2)**: 109-125.
- [12] X. Chen, S.T. Cheung, S. So, S.T. Fan, C. Barry, J. Higgins, K-M. Lai, J. Ji, S. Dudoit, I.O.L Ng, M. van de Rijn, D. Botstein, P.O. Brown. Gene expression patterns in human liver cancers. *Molecular Biology of the Cell* 2002; **13**: 1929-1939.
- [13] G. Gamberoni, S. Storari, S. Volinia. Finding biological process modifications in cancer tissues by mining gene expression correlations. *BMC Bioinformatics* 2006; **7**: 6-15.
- [14] J.M. Berg, J.L. Tymoczko, L. Stryer. *Biochemistry*. New York, WH Freeman. 2002.
- [15] M.H. Glickman, A. Ciechanover. The ubiquitin-proteasome pathway: destruction for the sake of construction. *Physiology Review* 2002; **82**: 373-428.

Table 1. Participants' post-task comments

Group: Bipartite graph	
Participants	Comments
1	Visually easy to look at but the line crossings may be a bit confusing if the number of connections increases. It is a good visualization tool for array data.
2	The fonts of the GO terms are difficult to read. Recommend colour-coded lines and text for each GO term. I would like to have a 3D spherical arrangement of the correlated gene clusters to complement the 2D visualization so that I can interpret the data for tasks I and J.
3	The visualization is helpful because it does give you some idea of gene co-expression in the context of GO terms. However, when the relationship between various GO terms becomes more complex, the appeal of the visualization diminishes. The visualization is not a great deal relevant to my research because GO terms describe a lot of generic processes and bear very little relationship to the actual pathophysiological data described in the literature.
4	Spontaneously, I would say that the GO clusters are not very accessible. It requires a certain amount of dedication to understand them. We use microarrays in several ways to understand gene regulation on the post-transcriptional level. Several projects have come up with gene lists in which GO terms are significantly enriched.
5	The 'normal' diagram was more difficult to read than the 'disease' diagram because of the increase in the number of lines causing congestion in the diagram. It would be good to have a function where clicking on a gene cluster changes the colour of the lines to the GO terms. I haven't used a data visualisation program on my data yet, so this looks good and will be helpful.
6	Summarize the results nicely by showing all the relationships between gene groups and the related GO terms. However when there are a lot of interactions between groups it's a little hard to read the results.
7	The lines were a bit vague when interpreting the correlations but the visualization could be relevant in the future for interpreting data.
Group: Block matrix	
8	These GO clusters are quite easy to understand and analyse. It can get a bit tricky when looking for a gene ontology that comes up not only in one cluster but multiple clusters. If the dataset of the normal tissue is bigger than what has been shown, it will be even more difficult to do task D. Yes, it is relevant. I need to use gene ontology annotation in all my microarray experiments. At present, I have to compare the gene ontology terms between two datasets by visually inspecting them on Excel spreadsheets.
9	It was difficult to relate GO terms that were represented more than once in the diagram. The visualization is very relevant. We are doing a lot of microarray work and have large datasets which we would like to relate to functional outcomes.
10	A bit confusing in regards to understanding the difference between co-expressed and co-regulated, and being able to compare disease to normal tissue using the cluster patterns. However, it provided a good way to visually inspect the groups of genes regulated in each tissue, as well as between tissue types, I think a spreadsheet would also have to be provided for the comparison between disease and normal.
11	I understood the concepts behind the visualisation graphics. It was quite easy to follow and then draw conclusions from what is being presented.
12	Easy to visualise. Perhaps might even be better if each of the bioprocesses had a different colour code.
13	A few questions were difficult without any experience in microarrays.
14	The graphs were easy enough to understand for a person who has no previous experience with gene arrays.

APPENDIX

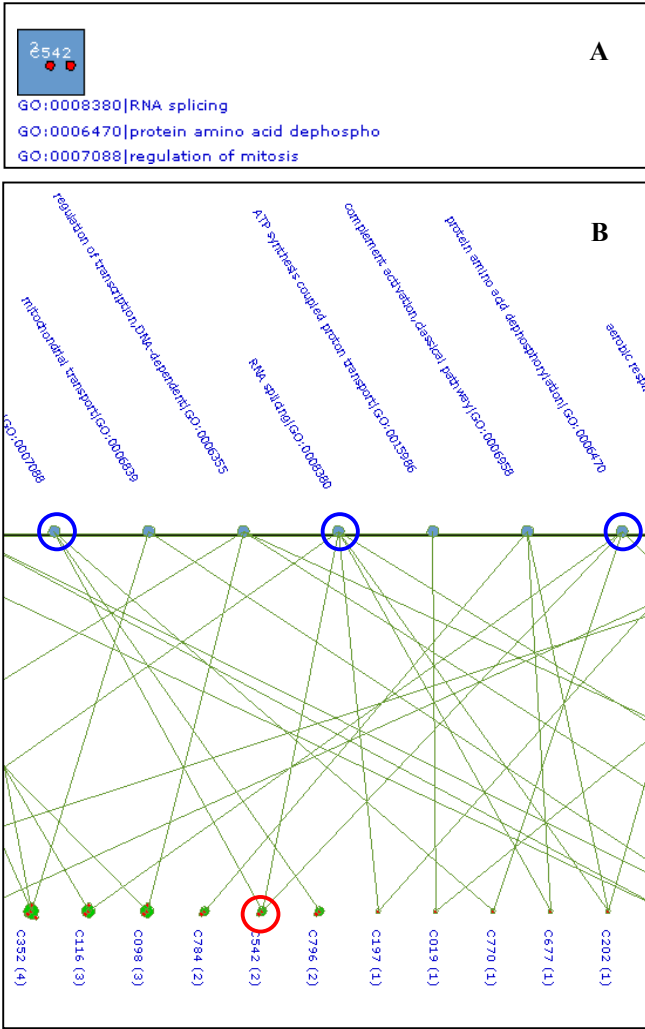


Figure S1. Visualization of cluster C542 in two representations. (A) Block matrix; (B) Bipartite graph. Blue circles from left to right are GO:0007088, GO:0008380, and GO:0006470. C542 is circled red.

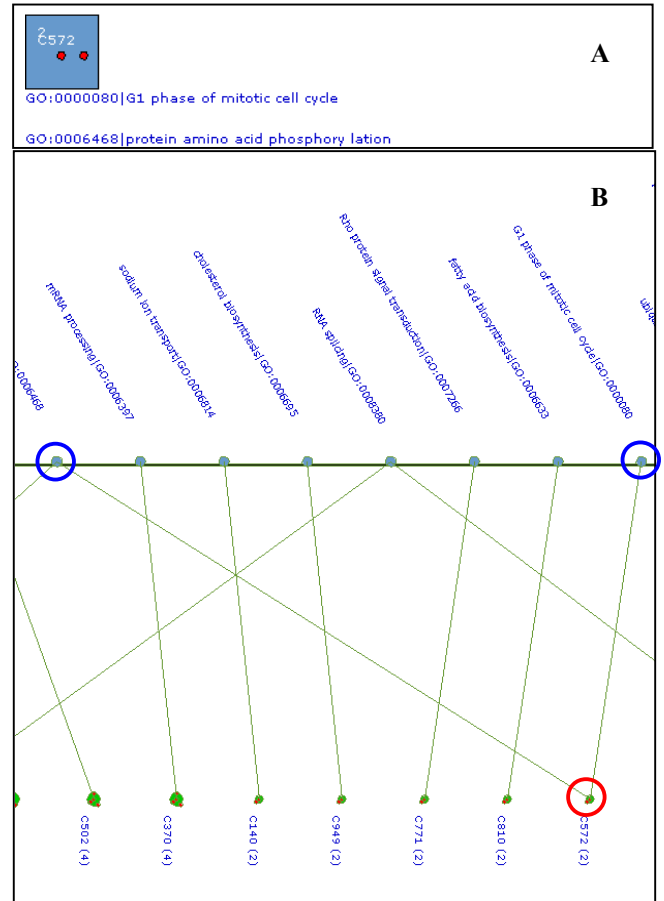


Figure S2. Visualization of cluster C572 in two representations. (A) Block matrix; (B) Bipartite graph. Blue circles from left to right are GO:0006468 and GO:0000080. C572 is circled red.

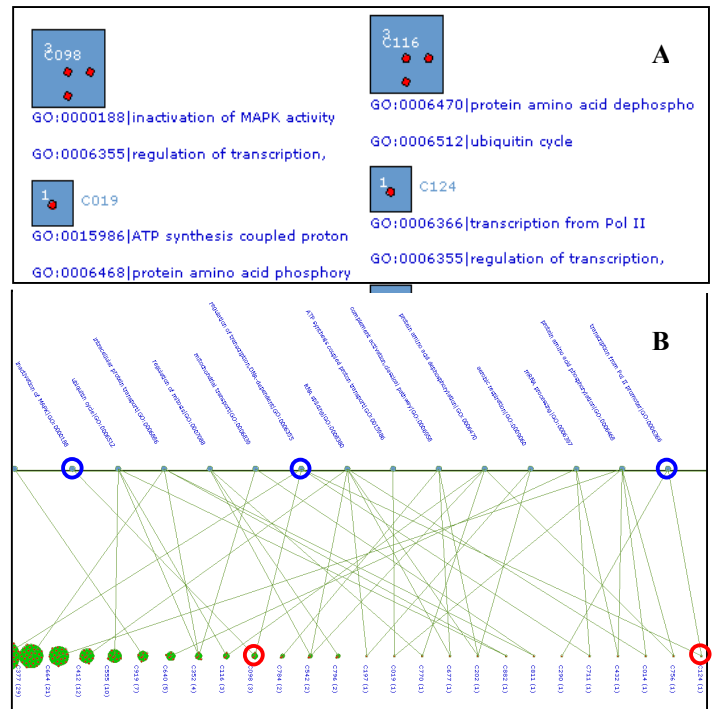


Figure S3. Visualization of clusters C098 and C124 in two representations. (A) Block matrix; (B) Bipartite graph. Blue circles from left to right are GO:0000188, GO:0006355, and GO:0006366. Red circles from left to right are C098 and C124.

School of Information Technologies, J12
The University of Sydney
NSW 2006 AUSTRALIA
T +61 2 9351 3423
F +61 2 9351 3838
www.it.usyd.edu.au

ISBN 978-1-74210-052-4