Extracting and Explaining Biological Knowledge in Microarray Data

Paul J. Kennedy[1], Simeon J. Simoff[1], David Skillicorn[2], and Daniel Catchpoole[3]

[1] {paulk,simeon}@it.uts.edu.au
Faculty of Information Technology, University of Technology, Sydney, PO Box 123,
Broadway, NSW 2007, AUSTRALIA
[2] skill@cs.queensu.ca
School of Computing, Queen's University, Kingston, Ontario, CANADA
[3] DanielC@chw.edu.au
The Oncology Research Unit, The Children's Hospital at Westmead,
Locked Bag 4001, Westmead NSW 2145, AUSTRALIA

Keywords: cluster analysis, bioinformatics, cDNA microarray, gene ontology, matrix decomposition methods.

Abstract. High throughput technologies produce large biological datasets that may lead to greater understanding of the biological mechanisms behind diseases such as cancer. However, progress has been slow in extracting meaningful information from these datasets. We describe a method of clustering lists of genes mined from a microarray dataset using functional information from the Gene Ontology. The method uses relationships between terms in the ontology both to build clusters and to extract meaningful cluster descriptions. The approach is general and may be applied to assist explanation other datasets associated with ontologies.

1 Introduction

Rapid developments in bio–technology, measurement and collection of diverse biological and clinical data have led to revolutionary changes in bio–medicine and biomedical research. The data collected in bio–medical experiments or as a result of medical examination ranges from gene expression levels measured using microarray technologies to data collected in therapy research. Researchers are looking at discovering relations between patterns of genes (sequences, interactions between specific genes, dependencies between changes in gene expressions and patient's responses to treatment). The confluence of bio–technology and statistical analysis is known as bioinformatics. The "classical" statistical techniques used in bioinformatics — a broad range of cluster, classification and multivariate analysis methods, have been challenged by the large number of genes that are analysed simultaneously and the curse of dimensionality of gene expression measurements. As a rule, the gene–to–data points ratio is high, the so-called "wide" data problem. In other words, if we are looking at N genes (collected as a result of identical microarray measurements) and our sample is of size m

(corresponding to samples from patients), then usually $N \gg m$ (in other words we are looking typically at tens of thousands of genes and only tens to hundreds of patients)). There is a number of ways in which data mining is expected to be able to assist the bio–data analysis (see [1] for brief overview).

One important area are the tasks of similarity search, comparison and grouping of gene patterns and assisting in understanding these patterns in medical bio–data, as many diseases are triggered by a combination of genes acting together. The diagram presenting the proposed methodology is shown in Fig. 1. The methodology includes three stages. Stage 1 ("DM1: extract") is a data–driven data mining cycle, during which the aim is to reduce the vast number of genes coming out of the microarray experiments to dozens of genes. The matrix decomposition methods that are used are a way of transforming microarray datasets into new forms that reveal their structure more clearly. Two different techniques are used: singular value decomposition (SVD) and semidiscrete decomposition (SDD). The output of this stage is interesting from a statistical point of view (an example of visualised clusters is shown in Fig. 2), however it is difficult for biological interpretation.

Stage 2 ("DM2: explain") aims at assisting the interpretation of the outputs of Stage 1. The results of the Stage 1 of our proposed methodology give a list of genes associated with high risk patients. Although such a list is important information to biologists, it is of limited use and appeal. This is because biologists wish to gain an understanding of the disease. They ask questions like: why are these genes important? and what is the biological meaning of the genes?

During the next stage of our methodology, data mining is applied to assist the explanation of the relations between the genes in the clusters identified in Stage 1. The data mining algorithms operate over the information from the Gene Ontology [2] dubbed "Congo". The list of genes is reclustered into groups of genes with similar biological functionality. Descriptions of the clusters are automatically determined using the Gene Ontology data and provided to biologists for interpretation which leads to the Stage 3 ("DM3: generate hypotheses").

Stage 3 aims to summarise what is known about the genes and to coarsely group them in the context of the microarray measurements, so that biologists can quickly focus their energies on potentially promising hypotheses. Depending on the formulated hypotheses biologists and data miners may return to Stage 1.

The focus of this paper is on Stage 2. In the next section, we present the data that is used in our methodology.

2 The Data

The broad goals of this work are to improve the understanding of genes related to a specific form of childhood cancer. Three forms of data are combined at different stages. Patient data includes data about the relative expression levels of genes combined with clinical data concerning the tumours and patients. Based on this data during the Stage 1 we produce a list of genes that are associated with high risk patients. In the second step, ontological terms (from the Gene Ontology data) associated with the genes are used to cluster the genes into functional groups.

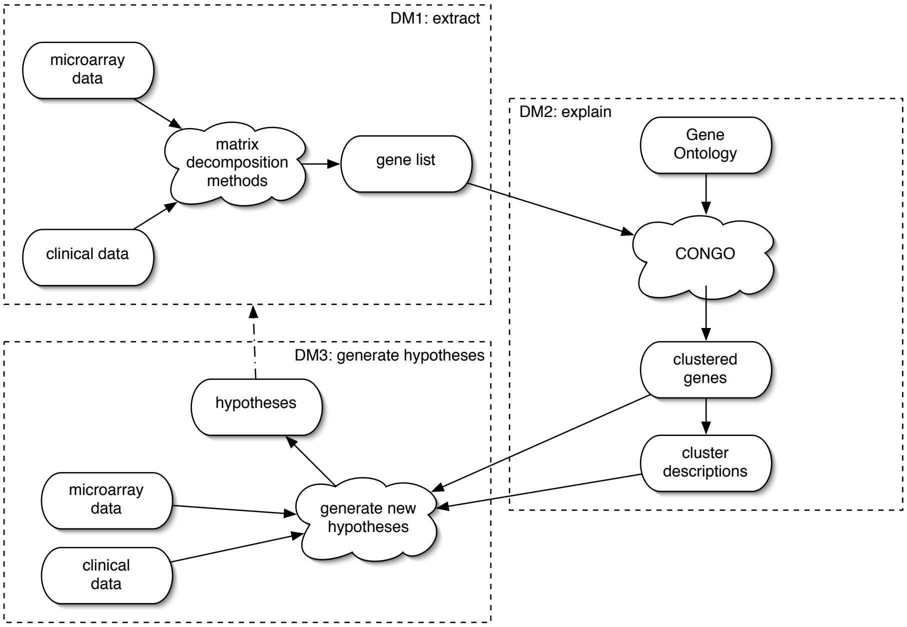

Fig. 1. Diagram showing methodology used to analyse microarray data

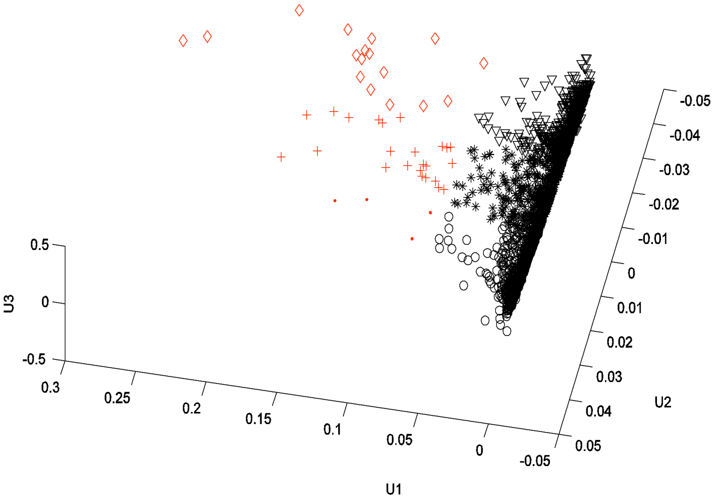

Fig. 2. Diagram showing an example of the output of DM1

Patient Data

Microarray and clinical data are available for nine patients.

The cDNA microarray experiment is a recent technology [3] available to cellular biologists that measures the relative expression levels of thousands of genes in cells at one time. Microarrays are microscope slides with a grid of DNA probes (called *spots*), each associated with a different gene. A sample of complementary DNA (cDNA) is taken from cells (indicative of the expression of genes at a time) and labelled with a fluorescent dye (green). It is mixed with another control sample of cDNA labelled with a different coloured dye (red). The mixture is washed over the microarray and cDNA molecules hybridize with the matching DNA probe on the slide. Measurements of the fluorescence of the spots under laser light are taken and normalised. The resulting gene expression data is a *set of log ratios* for each gene on the microarray. Usually between 2 and 10 repeat experiments of the same data (ie. patient) are made. For each patient, there are around 9000 genes with between 2 and 10 log ratios (ie. experiment repeats) for each gene.

A *small clinical dataset* is associated with each patient's cDNA microarray dataset that describes the patient in detail, as well as the effect of different treatment protocols. Our method uses only one piece of information from this dataset: a binary indicator of whether the patient is classified as high risk. Of the nine patients, four are labelled as high risk.

Gene Ontology Data

The Gene Ontology [2] is a large collaborative public set of controlled vocabularies constructed by researchers world–wide. Gene products are described in terms of their effect and known place in the cell. Terms in the ontology are interrelated. For example, a "glucose metabolism" *is a* "hexose metabolism" (see Fig. 3). Each node is a term in the ontology. Inside each node is the identifier for the term and beside is the term itself. More general terms are at the top of the diagram. All links shown are is–a relationships that are directed upwards. There are currently around 16,000 terms in the Gene Ontology and each gene is associated with between two and ten terms.

Each term in the ontology has a number of attributes: the term itself (eg. glycolysis), a unique accession number (eg. GO:0006096), and a definition (eg. the breakdown of a monosaccharide (generally glucose) into simpler components, including pyruvate). There may also be technical references to the definition (eg. links to PubMed articles), cross references into other biological databases, synonyms and comments.

There are a number of benefits of using the Gene Ontology as part of the data mining process. It is large (7045 terms in the Molecular Function ontology, 7763 terms in the Biological Process ontology and 1335 terms in the Cellular Component ontology as of 16 September 2003 [4]) and well worked on by researchers (16 member organisations of the Gene Ontology Consortium as of August 2003 [4]). Entries are curated before being added to the ontology. The ontology may

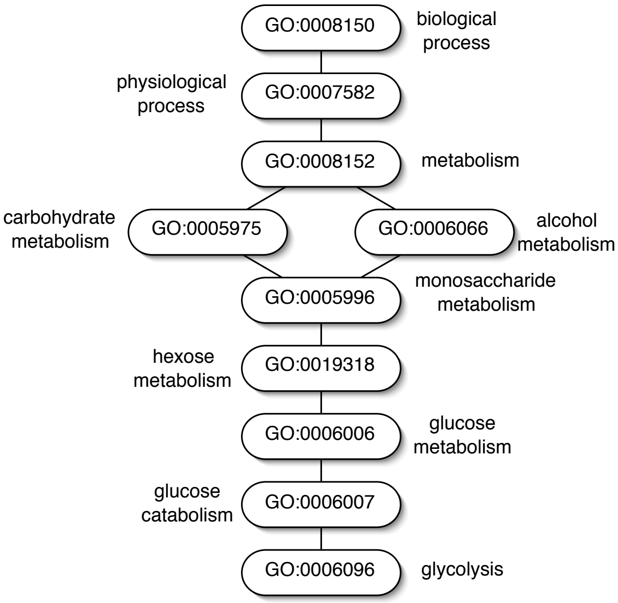

Fig. 3. Diagram showing an example of GO hierarchy from the "biological processes" ontology

be accessed in the RDF XML file format. In this computer legible form it is easier to apply the information to data mining methods and immediately richer than by determining similar information with text mining methods.

GO terms may be associated with genes using databases like SOURCE [5] as long as accession numbers of genes or gene names are known. This is true for the vast majority of the genes in our microarray dataset. See table 1 for an example.

3 DM2: Assisting Biological Explanation

The cluster analysis and visualisation described in this paper takes as input (i) a list of genes highlighted from a previous data mining step and (ii) data from the Gene Ontology. The previous data mining step used gene expression data (from cDNA microarray experiments) and clinical data describing the tumour cells in detail, effect of drug protocols and (human) classifications of patients into high or low risk categories. cDNA microarray experiments are a recent technology available to cellular biologists that measure the relative expression levels of thousands of genes in cells at one instant. Expression levels of genes in a test sample (i.e. tumour cells) compared to genes in a control sample (i.e. "normal" cells) are measured.

Table 1. GO terms associated with an example gene (named CLK1) for each of the three ontologies.

CLK1 (CDC–like kinase 1)
Molecular Function
GO:0004715 non–membrane spanning protein tyrosine kinase activity GO:0005524 ATP binding activity GO:0004674 protein serine/threonine kinase activity GO:0016740 transferase activity
Biological Process
GO:0006468 protein amino acid phosphorylation GO:0008283 cell proliferation GO:0000074 regulation of cell cycle
Cellular Component
GO:0005634 nucleus

Gene Ontology terms are associated with each gene in the list by searching in the SOURCE database [5]. The list of genes is clustered into groups with similar functionality using a distance measure that explicitly considers the relationship between terms in the ontology. Finally, descriptions of each cluster are found by examining Gene Ontology terms that are representative of the cluster. Graphs of Gene Ontology terms for each cluster together with cluster descriptions give a visualisation of each cluster in functional terms.

Taking the list of genes associated with high risk patients identified in Stage 1 (an example of such genes are shown in the first column in Table 2), we reclustered them using terms in the Gene Ontology (the GO:*nnnnnnn* labels in the right column in Table 2) into groups of similarly described genes, for example, genes that control signal transduction, or genes associated with transcription regulatory behaviour.

Currently, biologists would take such a list of genes and search one–by–one through Internet databases and search engines comparing the massive amount of information about each gene in an effort to find commonalities and differences. The aim of this step in our methodology is to summarise what is known about the genes and coarsely group them so that biologists can quickly focus their energies into promising areas.

Clustering data according to an ontology is a new procedure described in [6]. It entails using a special distance measure that considers the relative positions of terms in the ontological hierarchy.

The distance measure used essentially compares the number of GO terms that two data points have in common to the total number of GO terms associated with the data points. Since terms higher in the ontology are also important to the comparison they are also included but are “weighted down” and count for less than the lower level terms. We calculate “weighted” cardinalities of the bags of GO terms in common (the intersection) between data points and in total (the

Table 2. The first few rows of the dataset for the second step in the methodology.

Gene	GO terms directly associated with gene
AA040427	GO:0004715 GO:0005524 GO:0004674 GO:0006468 GO:0008283 GO:0000074 GO:0005634 GO:0016740
AA046690	GO:0003777 GO:0005524 GO:0007018 GO:0005871
AA055946	GO:0004894 GO:0005057 GO:0004888 GO:0007166 GO:0006968 GO:0005887

union). Terms then are weighted by their distance from the GO terms directly associated with the genes.

As described in [6] the distance measure used is an extension of the Tanimoto similarity measure between sets [7] [8]. The distance measure used is

$$D_{X,Y} = 1 - \frac{n'_{X \cap Y}}{n'_X + n'_Y - n'_{X \cap Y}} = 1 - \frac{n'_{X \cap Y}}{n'_{X \cup Y}} \tag{1}$$

where X and Y are the two bags of GO terms being compared and n'_X, n'_Y and $n'_{X \cap Y}$ are the weighted cardinalities of the bags X, Y and $X \cap Y$ respectively given by

$$n'_X = \sum_{i \in X} c^{d_i} \tag{2}$$

where X is the bag of GO terms, d_i is the distance of element of X with index i from its associated descendent in the original set of GO terms for the gene, and c is the weight constant. The weighted cardinality of the other bags is similarly defined.

The more general terms (ie. those higher in the ontology) provide a context for the more specific terms directly associated with genes. The c parameter allows variation of the importance of the "context" to the comparison. A value of $c = 0$ means that higher level terms are not considered, whilst a value of 1 considers all terms equally. Plainly, the latter is unreasonable because very general terms would be given undue importance. We arbitrarily chose $c = 0.9$ for our experiments.

The particular clustering algorithm is not as important as the distance measure. We used the Modified Basic Sequential Algorithmic Scheme (MBSAS) as described in [7] as it was simple and did not a priori presume a fixed number of clusters. The details of the clustering algorithm are presented in [6]. Table 3 shows the algorithm parameters necessary for understanding the experimental results in the next section.

4 Results of DM2

With the parameters values given above five clusters are found as shown in Table 4. Half of the genes have been allocated to one cluster. The rest of the

Table 3. Parameters used in the Modified Basic Sequential Algorithmic Scheme clustering algorithm. The last parameter is used only in the distance measure and is not formally part of MBSAS. See text for a detailed description of c.

Parameter	Meaning	Value
Θ	Minimum distance for points to be considered to be in the same cluster. (Theodoridis and Koutroumbas [7] call this the "threshold of dissimilarity").	0.001
q	Maximum allowable number of clusters.	0.1
M_1	Minimum distance for clusters to be deemed separate before they are merged.	5
c	Discount weight applied to GO nodes in the ontology.	0.9

genes have been split into four smaller clusters with one cluster containing only two genes. Such a tabular representation does not increase our understanding of the clusters as the gene accession codes are not descriptive. Moreover, it does not help biologists.

Table 4. Clusters found with the MBSAS clustering algorithm. The codes AA*nnnn* are GenBank accession codes.

Cluster Number	Gene Count	Genes
0	6	AA040427 AA406485 AA434408 AA487466 AA609609 AA609759
1	2	AA046690 AA644679
2	6	AA055946 AA398011 AA458965 AA487426 AA490846 AA504272
3	9	AA112660 AA397823 AA443547 AA447618 AA455300 AA478436 AA608514 AA669758 AA683085
4	20	AA126911 AA133577 AA400973 AA464034 AA464743 AA486531 AA488346 AA488626 AA497029 AA629641 AA629719 AA629808 AA664241 AA664284 AA668301 AA669359 AA683050 AA700005 AA700688 AA775874

With this in mind, we plotted the subset of terms associated with the clustered genes as nodes on a graph with relationships represented by edges and the GO nodes of a cluster localised to one part of the graph as much as possible (Fig. 4). The clusters are represented by the five large boxes with the cluster numbers (as listed in Table 4) given inside each box. Nodes inside the clusters are the GO terms associated with genes in that cluster. More general terms are on the right hand side of the diagram. Edges between nodes represent the links in the ontology. Some terms, particularly the more general ones at the right hand side of the diagram, have links from terms in a different cluster. Each node is shown in only one cluster box, but links between the boxes show where GO terms are shared by genes in the different clusters. The grey scale of the link represents the cluster that link is in. Also, a darker grey scale is used for links in the original dataset whilst a lighter shade is used for relationships inferred from traversing the ontology.

The GO terms lying along the right edges of the cluster boxes (particularly in cluster 1) are important. These terms are part of the most general descriptions for a cluster that *do not also* describe another cluster. Figure 5 shows a closer view of the terms at the right edge of cluster 1. These terms are used to automatically determine cluster descriptions. They provide a functional description of a cluster. Starting with all the GO terms directly associated with genes in a particular cluster, we climb the hierarchy replacing GO terms with their parent terms. Terms are replaced only if the parent node is *not* associated with genes in another cluster.

Cluster descriptions derived in this way are shown in Table 5. Only the *is–a* relationships were followed to build this table. There are far fewer *part–of* relationships in the hierarchies so we do not believe that omitting them affects the results. The terms listed in the table are associated only with genes in each cluster and not in any other cluster. Cluster 0 in Table 5 has no terms that are associated with more than one gene. This suggests that the genes in the cluster are either unrelated or related only in ways that are sufficiently high level that the terms exist in other clusters. This suggests that the quality of the cluster is not good. Cluster 1 contains at least two genes that are related to the cell cytoskeleton and to microtubules (microtubules are components of the cytoskeleton). Cluster 2 contains three or four genes associated with signal transduction and cell signalling. Cluster 3 contains three or four genes related to transcription of genes and cluster 4 seems to contain genes associated with RNA binding.

5 Conclusions

This paper presents a methodology for extracting and explaining biological knowledge from microarray data. This is a two step approach involving clustering lists of genes mined from a microarray dataset using functional information from the Gene Ontology. The method uses relationships between terms in the ontology both to build clusters and to extract meaningful cluster descriptions. In this work we have limited generation of cluster descriptions to the “is-a” ontological relationships.

Applying information from the Gene Ontology to cluster genes allows for an understanding of the genes and their interrelationships in functional terms. Currently biologists search through such lists gene–by–gene analysing each one individually and trying to piece together the many strands of information. Automating the process, at least to some extent, allows biologists to concentrate more on the important relationships rather than the minutiae of searching. Consequently they are enabled to formulate hypotheses to test in future experiments.

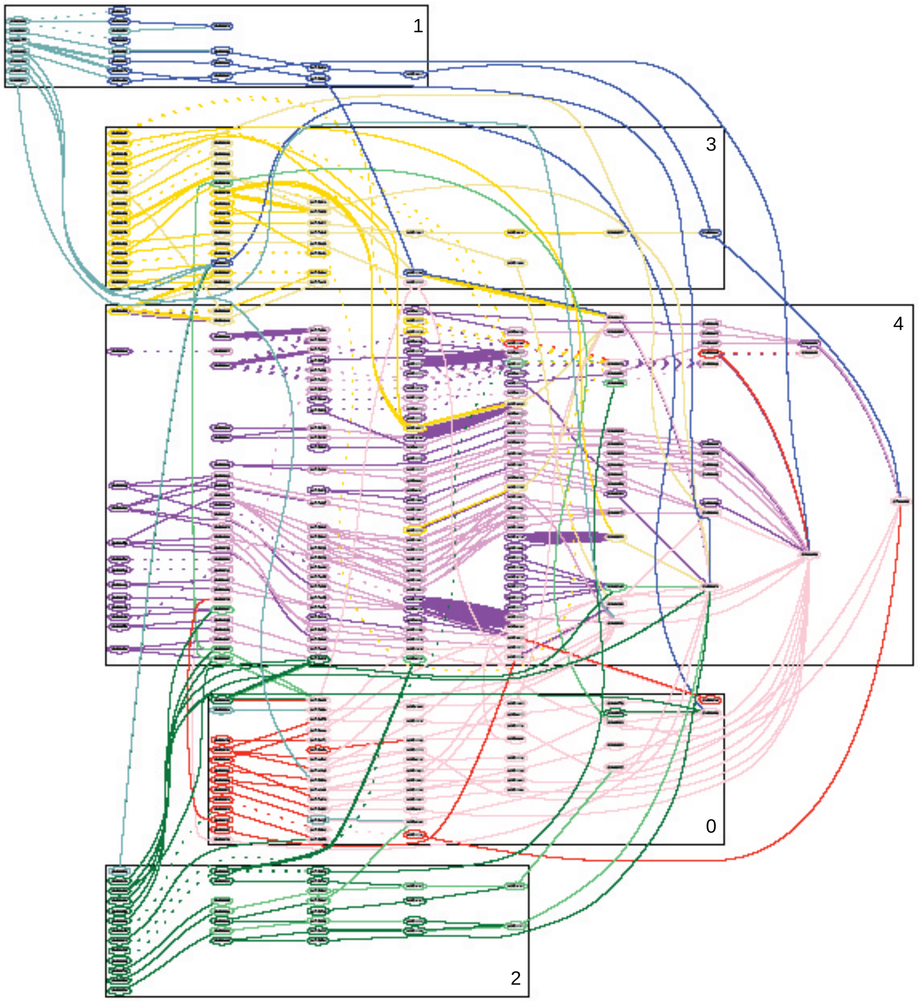

Fig. 4. Diagram showing parts of the GO hierarchy associated with genes being clustered. More general terms are at the right of the diagram. See text for description of graph.

Table 5. Principal cluster descriptions for the genes clustered with the MBSAS algorithm derived as stated in the text. The last column gives the number of genes in the cluster associated with the term.

GO ID	GO Term	Number of Genes
Cluster 0 — 6 genes		
	20 GO terms but each associated with only one gene	1
Cluster 1 — 2 genes		
GO:0008092	cytoskeletal protein binding activity	2
GO:0007028	cytoplasm organization and biogenesis	2
GO:0003774	motor activity	2
GO:0005875	microtubule associated complex	2
	5 GO terms but each associated with only one gene	1
Cluster 2 — 6 genes		
GO:0004871	signal transducer activity	4
GO:0007154	cell communication	4
GO:0005887	integral to plasma membrane	3
GO:0005886	plasma membrane	3
GO:0005194	cell adhesion molecule activity	2
	11 GO terms but each associated with only one gene	1
Cluster 3 — 9 genes		
GO:0030528	transcription regulator activity	4
GO:0008134	transcription factor binding activity	3
GO:0006366	transcription from Pol II promoter	3
GO:0003700	transcription factor activity	3
GO:0006357	regulation of transcription from Pol II promoter	3
	5 GO terms but each associated with only two genes each	2
	13 GO terms but each associated with only one gene	1
Cluster 4 — 20 genes		
GO:0003723	RNA binding activity	10
GO:0030529	ribonucleoprotein complex	9
GO:0009059	macromolecule biosynthesis	9
GO:0006412	protein biosynthesis	9
GO:0005829	cytosol	9
GO:0003735	structural constituent of ribosome	8
	2 GO terms but each associated with only four genes each	4
	5 GO terms but each associated with only three genes each	3
	1 GO term associated with only two genes	2
	33 GO terms but each associated with only one gene	1

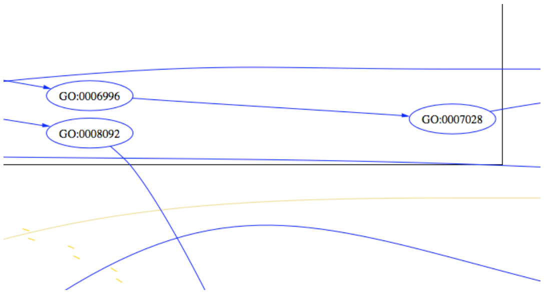

Fig. 5. Diagram showing a close up of the most general GO terms in the large cluster. See text for further description.

The approach is general and may be applied to assist explanation other datasets associated with ontologies.

Acknowledgements

We would like to thank the University of Technology, Sydney and The Children's Hospital at Westmead for supporting this project.

References

1. Han, J.: How can data mining help bio–data analysis. In: Proceedings 2nd Workshop on Data Mining in Bioinformatics BIOKDD02, in conjunction with ACM SIGKDD 8th International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, ACM Press (2002)
2. Ontology Consortium, T.G.: Gene Ontology: tool for the unification of biology. Nature Genetics **25** (2000) 25–29 PubMed ID:10802651.
3. Baldi, P., Hatfield, G.W.: DNA microarrays and gene expression. Cambridge University Press, Cambridge, UK (2002)
4. Ontology Consortium, G.: Gene Ontology Consortium. Available on: `http://www.geneontology.org` (2003) Viewed at 15 October 2003.
5. Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J.C., Hernandez-Boussard, T., Rees, C.A., Cherry, J., Botstein, D., Brown, P.O., Alizadeh, A.A.: SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Research **31** (2003) 219–223
6. Kennedy, P.J., Simoff, S.J.: CONGO: Clustering on the Gene Ontology. In: Proceedings 2nd Australasian Data Mining Workshop ADM03, in conjunction with Congress on Evolutionary Computation, Canberra, Australia, University of Technology, Sydney (2003)
7. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press, San Diego, USA (1999)
8. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Second edn. John Wiley and Sons, New York (2001)