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.

Abstract. This paper describes 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 of other datasets associated with ontologies.

1 Introduction

Rapid developments in measurement and collection of diverse biological and clinical data offer researchers new opportunities for discovering relations between patterns of genes. The "classical" statistical techniques used in bioinformatics have been challenged by the large number of genes that are analysed simultaneously and the curse of dimensionality of gene expression measurements (in other words we are looking typically at tens of thousands of genes and only tens of patients). Data mining is expected to be able to assist the bio–data analysis (see [1] for brief overview).

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 include cDNA microarray and clinical data for 9 patients. 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. Clinical data describe a patient in detail, as well as the effect of different treatment protocols. Of the nine patients, 4 are labelled as high risk.

The task is to assist in understanding gene patterns in such biodata. Proposed methodology is shown in Fig. 1. It includes 3 stages. Stage 1 ("DM1: extract") is a data mining cycle, which reduces the vast number of genes coming

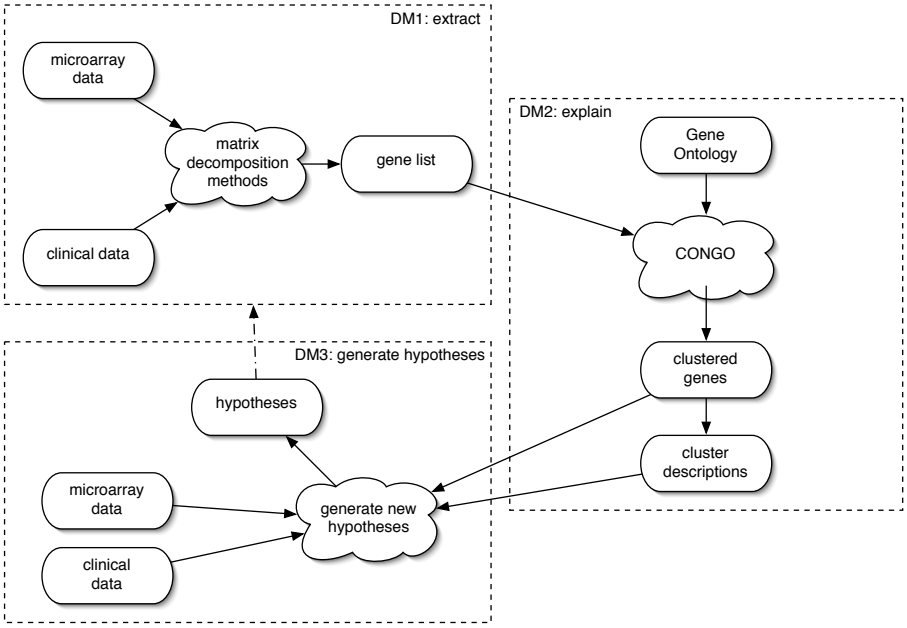

Fig. 1. Diagram showing methodology used to analyse microarray data

from the microarray experiments to dozens of genes. Techniques used are described in detail in [2]. The output of this stage is interesting from a statistical point of view, however it is difficult for biological interpretation. Stage 2 ("DM2: explain") aims at assisting the interpretation of these outputs. The list of genes is reclustered over a gene ontology [3] into groups of genes with similar biological functionality. Descriptions of the clusters are automatically determined for biological interpretation. Stage 3 ("DM3: generate hypotheses") aims to summarise what is known about the genes and to group them in the context of the microarray measurements. Biologists then can formulate potentially promising hypotheses and may return to Stage 1.

2 DM2: Assisting Biological Explanation

The focus of this paper is on Stage 2. The cluster analysis and visualisation described in this paper takes as input (i) a list of genes highlighted from "DM1: extract" and (ii) data from the Gene Ontology. Clustering data according to an ontology is a new procedure described in [4]. It entails using a special distance measure that considers the relative positions of terms in the ontological hierarchy. The particular clustering algorithm is not as important as the distance measure. Details of the algorithm are presented in [4]. Recent work in [5] takes a similar approach. We use the Gene Ontology [3], a large collaborative public

Table 1. 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

Table 2. Discovered clusters. 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

set of controlled vocabularies, in our clustering experiments. Gene products are described in terms of their effect and known place in the cell. Terms in the ontology are interrelated: eg. a "glucose metabolism" *is a* "hexose metabolism". Gene Ontology terms are associated with each gene in the list by searching in the SOURCE database [6]. 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.

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 1), we reclustered them using terms in the Gene Ontology (the GO:*nnnnnnn* labels in the right column in Table 1) into groups of similarly described genes.

3 Results of DM2

Five clusters are found as shown in Table 2. Half of the genes have been allocated to one cluster. The rest of the genes have been split into four smaller clusters with one cluster containing only two genes.

Associated GO terms automatically determine functional descriptions of clusters. Starting with all the GO terms directly associated with genes in a particular

Table 3. Principal cluster descriptions for the genes. Last column is 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

cluster, we climb the ontology replacing GO terms with their parents. 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 3. 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 3 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 (ie. 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 contains genes associated with RNA binding.

4 Conclusions

We present a methodology for extracting and explaining biological knowledge from microarray data. Applying terms from the Gene Ontology brings an understanding of the genes and their interrelationships. 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. 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: Proc. 2nd Workshop on Data Mining in Bioinformatics BIOKDD02, ACM Press (2002)
2. Skillicorn, D., et al.: Strategies for winnowing microarray data. In: Proc. SIAM Data Mining Conf. (accepted 2004)
3. Ontology Consortium, T.G.: Gene Ontology: tool for the unification of biology. Nature Genetics **25** (2000) 25–29 PubMed ID:10802651.
4. Kennedy, P.J., Simoff, S.J.: CONGO: Clustering on the Gene Ontology. In: Proc. 2nd Australasian Data Mining Workshop ADM03, University of Technology, Sydney (2003)
5. Lee, S.G., et al.: A graph–theoretic modeling on GO space for biological interpretation of gene clusters. Bioinformatics **20** (2004) 381–388
6. Diehn, M., et al.: SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Research **31** (2003) 219–223