Gene Expression and Annotation

Clare Foyle[1], Nick Fieller[1], AstraZeneca R&D[2]

[1]Department of Probability and Statistics, University of Sheffield. [2]AstraZeneca R&D, Alderley Park, UK

Email: stp02cf@sheffield.ac.uk

Introduction

Statistical analysis of gene expression data obtained from microarrays often involves pre-selection of ‘potentially interesting’ genes and then further analysis to determine ‘interesting groups’ of genes that may behave similarly and play an important role in the biological processes involved in the condition of interest. In parallel there is a wealth of published information on the role of genes in various diseases, biological processes &c. This information is known collectively as ‘annotation’ and provides a further route for relating genes into groups. The aim of our study is to explore ways of combining annotation information with gene expression data to provide an enhanced understanding and insight into the condition under study.

Available Data

Gene expression data is available for 58 cancer patients, plus 5 healthy controls and samples from ‘smooth muscle’ and non-diseased muscle from study subjects. Annotation information was obtained by text mining on subsets of genes for links between genes and MeSH (**Me**dical **S**ubject **H**eadings maintained by the US National Library of Medicine) terms in categories C & G (Diseases & Biological Sciences).

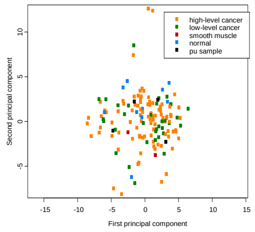
Selecting Probesets

Data are available from ~12,000 probesets analysed with Affymetrix GeneChips. Sets of 100 probesets were selected for further investigation. Selection was on the basis of high expression level (set A) & greatest variability over the known medico-demographic groups of subjects (Set B).

1

Exploratory Analysis

Figure 1: ***Principal Components 1 and 2 – 100 Randomly Selected Probesets***



Principal components analysis was carried out on the two selected datasets and for comparison on a random selection of probesets.

A cancer severity measure based on stage, grade and lymph node assessments subdivided subjects into ‘high level & ‘low level’ (or unknown if no information was available).

Figure 1 shows the PCA score plot on the first two components (14% & 8% of variation) for the randomly selected probesets.

There is little separation between the various severity measures and this provides a baseline for comparison with similar displays using probesets selected as ‘potentially interesting’ (i.e. highest expressing or most differentially expressed over subgroups). These are displayed in Figures 2 and 3.

2

PCA Plots

Figure 2: ***Principal Components 1 and 2 – 100 highest expressing probesets***

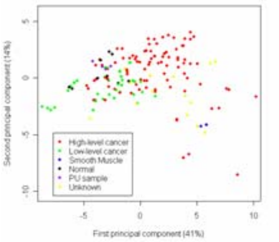
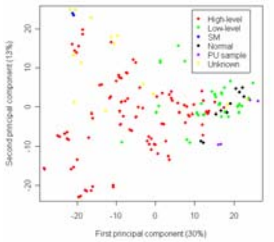


Figure 3: ***Principal Components 1 and 2- 100 most variable probesets***



Figures 2 and 3 show the score plots for the first two components on datasets A and B. These contain 55% and 43% of the variation. Clear structure is visible, especially on the first component, confirming that these selected subsets contain more relevant information than just random selection of probesets and, more pertinently, serve as the next baseline for comparison with enhancements achieved by incorporation of MeSH information.

3

Annotation Information

Annotation information includes qualitative data on specific genes that is available in the public domain. Of particular interest in this study are diseases and biological functions that have been linked to specific genes by co-citation in published literature. Pairs of genes that have been linked in this way to the same disease or biological function may be considered ‘related’. Disease and biological terms were taken from the MeSH listings and co-occurrences of these with selected genes determined by text mining. This results in a co-occurrence matrix of genes × MeSH terms.

MeSH Listings

The MeSH vocabulary is maintained and updated by the National Library of Medicine. It is organised in a hierarchical structure, starting with 15 root nodes. At each level of the 11-layer graph the terms become more detailed and specific. MeSH descriptors (i.e. diseases or biological functions) that are directly connected in the graph are closely related (mother/daughter pairs). Descriptors which are connected via a common node or parent are less closely related (siblings). Similarly, those connected by two nodes are one degree less closely related.

Two genes which are individually linked to MeSH descriptors which are distinct but are related by the MeSH DAG may be considered more distantly related than those linked to the same MeSH descriptor.

In total there are 22,997 descriptors in MeSH. A complication in the study is that a descriptor can occur several times within the hierarchical structure in different contexts.

4

The MeSH Data

MeSH terms are arranged in a hierarchy (DAG). A screen-grab is shown, left, of a sample MeSH descriptor from the website. It’s position in the DAG is described, it’s base node is F01.

If the term occurs more than once, more tree numbers are displayed here.

Finally, the unique identifier is given for each term.

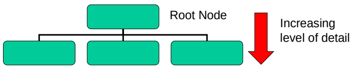
The example on the right shows a disease term with 3 positions on the DAG. The website also gives a representation of the DAG at each of the points in which it occurs.

Terms on the same level of the DAG are equally justified. Child nodes are shown for the node queried. Other nodes with children are indicated by a +. Terms close together on the DAG are more closely related than those far apart.

5

The Graph Structure

Figure 4: ***Annotation Graph Structure***



Nodes are annotation terms. Edges link nodes if the terms are related, becoming more informative and with more detail at each level of the graph.

The annotation information, *A*, takes the form of an *n x p* matrix, with *n* probesets and *p* annotation terms. The matrix shows the number of associations found linking an annotation term. Converting this to a binary matrix yields the occurrence matrix:

1 represents a link between annotation and probeset
0 is shown if no link has been found

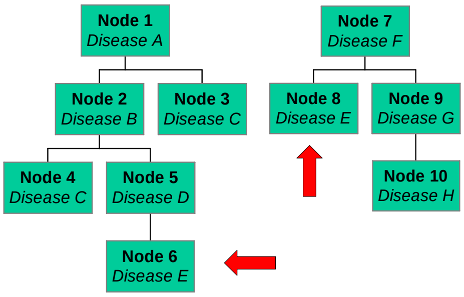
The square matrix AA^T gives the co-occurrence (or ‘similarity’) matrix of the probesets. Typically this matrix is large and sparse providing few links between genes of interest.

To provide further (but weaker) links between genes, the up-propagation of the MeSH DAG is used to establish relationships between genes. Initially, genes linked via mother-daughter pairs of descriptors can be determined. Repetition of the process allows successively weaker links to be determined via successively less closely related descriptors.

6

Up-Propagation

Figure 5: ***Node issues***



A complication in the up-propagation process is that descriptors can occur more than once in the hierarchy so some preliminary manipulation of the matrix is required. Figure 5 shows an example of one disease (or annotation term) occurring in two separate places in the graph structure.

The up-propagation matrix, *P*, is the *p x p* matrix of the DAG indicating which nodes are directly linked. The matrix APA^T gives the probesets which are related by links to one-generation related MeSH descriptors and AP^rA^T gives those linked via r-generation related descriptor.

7

Further Issues

Further issues arise because of the magnitude of the matrices. Annotation terms can be regarded as subsets of the original set of 22,997 descriptors – in this case the ‘disease’ terms and ‘biological function’ terms were considered. Although these contain approximately 4000 and 2000 descriptors respectively, many descriptors actually occur more than once in the graph structure – they map to more than one node.

In the case of the disease terms, over 10,000 nodes have to be considered which proved to cause some computational problems. One solution is to split the set of nodes into more manageable subsets.

A second issue is that typically the gene expression data is obtained in terms of probesets (with several probesets comprising one gene) but annotation information is available at the level of genes. Again, some preliminary manipulation is required to match the forms of information.

Once the links between probesets are established these can be used to enhance the analysis of the gene expression information, for example by treating the expression levels as group responses.

http://www.shef.ac.uk/paspgr/foyle/ http://www.shef.ac.uk/nickfieller/

http://www.astrazeneca.co.uk/

8