Storage and analysis of microarray data

Chih-hung Jen, Ioannis Michalopoulos, Archana Sharma-Oates, Iain Manfield, Phil Gilmartin, Noel Buckley, Phil Quirke and David Westhead

Introduction
The group has a growing interest in data from post genomic research, including microarray based measurements of gene expression, and, more recently, tissue microarrays. Work is collaborative with local experimental groups who generate data, and we are responsible for three aspects of these projects: (i) appropriate storage and archiving of data according to international standards, and efforts to advance these standards; (ii) data analysis using methods of multivariate statistics; and (iii) the use of private and public data to make predictions and motivate experimental verification or refutation.

Plant post genomics
Microarray analyses are being carried out to identify \textit{in vivo} targets of plant GATA transcription factors in \textit{Arabidopsis thaliana}. GATA transcription factors are Type IV zinc finger proteins, found in all eukaryotes. 29 \textit{Arabidopsis thaliana} GATA genes are cloned, all sharing a CX\textsubscript{2}CX\textsubscript{18}CX\textsubscript{2}C zinc finger domain. Their mammalian orthologs show specificity of the conserved promoter element GATAAGG. \textit{Arabidopsis thaliana} GATA-2 and GATA-4 are Phytochrome A regulated, as opposed to GATA-1.

Identifying coexpressed genes from the microarray data can be used to assign potential functions to new genes and help the discovery of transcriptional regulation networks. Currently, the coexpressed genes are usually analysed by many sophisticated clustering algorithms e.g. SOM, hierarchical clustering, k-means clustering. However, these clustering approaches usually depend on the distance cut-off value or arbitrary k value to group the genes, and these criteria do not really indicate the significance of the similarity within the clusters. Besides, they assign particular genes to only one cluster that may cause loss information where genes may have multiple biological roles or respond to different transcription factors.

In order to identify the \textit{in vivo} potential targets of \textit{Arabidopsis} GATA family transcription factors using microarray data and avoid the drawbacks of clustering algorithms, we propose a novel robust approach of assessing the significance of relationships in expression. We developed a new WWW-based \textit{Arabidopsis} Co-Expression Tool (ACT) for plant gene analysis, based on large \textit{Arabidopsis thaliana} public microarray data sets consisting of 322 Affymetrix arrays (ATH1) from 51 different experiments, obtained from the Nottingham Arabidopsis Stock Centre (NASC). The co-expression analysis tool allows users to identify genes whose expression patterns are correlated across selected experiments or the complete data set. The output is the Pearson correlation coefficient, or r-value, which is a scale-invariant measure of expression similarity, and this is accompanied by probability (p) and expect (E) values reflecting statistical significance against a background of random chance correlations. The E value is calculated as a product of the number of genes on the array and the p value. The correlation coefficient (r) is used to rank the genes in descending order of correlation with the driver. In addition to r, p and E values, the output includes Affymetrix probe ID, AGI code and current annotation for each gene. Genes with strongly correlated expression patterns are likely to be under similar transcription regulatory mechanisms, or involved in related biological processes. We illustrate the applications of the software by analysing genes encoding functionally related proteins, as well as pathways involved in plant responses to environmental stimuli. The resource is freely available at http://www.arabidopsis.leeds.ac.uk/.
Based on the r-value derived from the correlation analysis tool, we can reveal the GATA coexpressed genes with confidence. An example result of the top fifty genes coexpressed with GATA-1 is shown in Table 1 and co-correlation plot of GATA2 and GATA-4 is shown in Fig. 1.

### Table 1: Top fifty correlated genes to GATA-1 Arabidopsis gene.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Correlation</th>
<th>p-value</th>
<th>e-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis</td>
<td>Related to GATA-1 expression</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 1:** Co-correlation scatter plot of GATA-2 against GATA-4.

**Human cancer pathology**

Human tissue samples are obtained from the individuals in a clinical trial (both cancerous and normal tissues from the same person) for biochemical and histopathological analyses.
Depending on the nature of the trial these tissue samples undergo extensive characterisation using a number of high-throughput molecular biology techniques. The high-throughput techniques most commonly used in cancer research are cDNA microarrays, comparative genomic hybridisation arrays and tissue microarrays. The purpose of the cDNA microarray approach is to gain an insight into the expression levels of all the predicted genes in the human genome with the aim of identifying a set of genes related to a clinical outcome that may be either up or down regulated in tumour verses normal tissue. Comparative genomic hybridisation (CGH-arrays) arrays are used to study chromosomal instability at a genome level within tumour verses normal tissues. TMA is a technique that enables the analysis of a large cohort of clinical specimens in a single experiment thereby studying the molecular alterations (at the DNA, RNA, or protein level) in thousands of tissue specimens in parallel. The aim of cDNA microarray and CGH-array techniques are to either identify biomarkers that can be verified by TMA.

Our involvement in this research involves analysis of the cDNA microarray and CGH-array data using statistical approaches, and the development of storage and analysis software for tissue microarray experiments, and area where we are contributing to the development of international standards, and the integration of this data with MIAME compliant microarray databases.

TMAs are used in the laboratory to assess on a large-scale the diagnostic and therapeutic significance of various genes and proteins in colorectal tumour samples. A relational database has been designed and implemented in MySQL. The information stored in the database include TMA design constructs, tissue staining protocols, the results including images scanned from digital slide scanners and the pathology reports associated with each tumour sample. Additional information includes experiment authors, dates of each experiment, quality of cores on each TMA slide and the storage location of each TMA within the laboratory. This database is interfaced with the World Wide Web (WWW) thereby enabling users to query and assimilate their own data into the database.

Collaborators
Profs. P.M.Gilmartin and N. Buckley, Dr. P. Devlin (plant project), Prof. P. Quirke (human cancer pathology).

Funding
This work is funded by the BBSRC.