

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 *in vivo* targets of plant GATA transcription factors in *Arabidopsis thaliana*. GATA transcription factors are Type IV zinc finger proteins, found in all eukaryotes. 29 *Arabidopsis thaliana* GATA genes are cloned, all sharing a CX₂CX₁₈CX₂C zinc finger domain. Their mammalian orthologs show specificity of the conserved promoter element GATAAGG. *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 *in vivo* potential targets of *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 *Arabidopsis* Co-Expression Tool (ACT) for plant gene analysis, based on large *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.

256916_at_AT3G24050 GATA transcription factor 1 (AtGATA-1)						
Probe Set	r-value	p-value	e-value	GeneID	Annotation	
250274_at0.683184	1.4e-45	2.9e-41	AT5G13020	expressed protein		
253780_at0.649790	5.2e-40	1.1e-35	AT4G28400	protein phosphatase 2C (PP2C), putative		
256853_at0.645309	2.6e-39	5.7e-35	AT3G18640	hypothetical protein		
256183_at0.628998	7.2e-37	1.6e-32	AT1G51660	mitogen-activated protein kinase kinase (MAPKK), putative (MKK4)		
255095_at0.610592	2.8e-34	6.1e-30	AT4G08500	mitogen-activated protein kinase kinase, putative		
253204_at0.609395	4.0e-34	8.8e-30	AT4G34460	transducin / G-protein beta-subunit (AGB1)		
263274_at0.605274	1.4e-33	3.2e-29	AT2G11520	protein kinase family		
258979_at0.602097	3.8e-33	8.4e-29	AT3G09440	heat shock protein hsc70-3 (hsc70.3)		
252449_at0.595278	3.0e-32	6.5e-28	AT3G47060	FtsH protease, putative		
250994_at0.584791	6.3e-31	1.4e-26	AT5G02490	heat shock protein hsc70-2 (hsc70.2) (hsp70-2)		
246292_at0.583759	8.4e-31	1.8e-26	AT3G56860	RNA recognition motif (RRM) - containing protein		
246529_at0.582779	1.1e-30	2.4e-26	AT5G15730	serine/threonine protein kinase, putative		
248195_at0.582351	1.3e-30	2.8e-26	AT5G54110	VAMP (vesicle-associated membrane protein)-associated protein family		
267341_at0.580437	2.2e-30	4.7e-26	AT2G44200	expressed protein		
252670_at0.577386	5.1e-30	1.1e-25	AT3G44110	DnaJ protein AtJ3		
250350_at0.575414	8.8e-30	1.9e-25	AT5G12010	expressed protein		
248131_at0.572371	2.0e-29	4.5e-25	AT5G54830	expressed protein		
245986_at0.569217	4.8e-29	1.1e-24	AT5G13160	protein kinase family		
267009_at0.568045	6.6e-29	1.4e-24	AT2G39260	middle domain of eukaryotic initiation factor 4G (MIF4G) domain-containing protein		
256620_at0.566753	9.3e-29	2.0e-24	AT3G22170	far-red impaired response protein -related		
247054_at0.566549	9.9e-29	2.2e-24	AT5G66730	zinc finger protein		
257484_at0.565353	1.4e-28	3.0e-24	AT1G01650	expressed protein		
261743_s_at	0.564090	1.9e-28	4.2e-24	AT1G08420	protein serine/threonine phosphatase alpha -related	
249613_at0.563312	2.3e-28	5.1e-24	AT5G37380	DnaJ protein family		
252906_at0.561615	3.7e-28	8.0e-24	AT4G39640	gamma-glutamyltransferase -related		
257883_at0.561464	3.8e-28	8.4e-24	AT3G16940	calmodulin-binding protein		
246221_at0.560417	5.0e-28	1.1e-23	AT4G37120	step II splicing factor - like protein		
262408_at0.559422	6.5e-28	1.4e-23	AT1G34750	protein phosphatase 2C (PP2C), putative		
261520_at0.559319	6.7e-28	1.5e-23	AT1G71820	SEC6 protein		
266800_at0.558941	7.4e-28	1.6e-23	AT2G22880	hypothetical protein		
255605_at0.554410	2.4e-27	5.3e-23	AT4G01090	expressed protein		
252862_at0.553816	2.8e-27	6.2e-23	AT4G39830	L-ascorbate oxidase, putative		
249988_at0.553640	2.9e-27	6.4e-23	AT5G18310	expressed protein		
259202_at0.552617	3.8e-27	8.4e-23	AT3G09100	mRNA capping enzyme, RNA guanylyltransferase -related		
263457_at0.5511520	5.1e-27	1.1e-22	AT2G22300	ethylene-induced calmodulin-binding protein, putative		
259341_at0.551249	5.4e-27	1.2e-22	AT3G03740	expressed protein		
251861_at0.551161	5.6e-27	1.2e-22	AT3G54810	GATA zinc finger protein		
255280_at0.550891	5.9e-27	1.3e-22	AT4G04960	receptor lectin kinase, putative		
256185_at0.549733	8.0e-27	1.7e-22	AT1G51700	Dof zinc finger protein		
248268_at0.547811	1.3e-26	2.8e-22	AT5G53480	importin beta, putative		
247874_at0.547468	1.4e-26	3.1e-22	AT5G57710	101 kDa heat shock protein; HSP101-related protein		
262054_s_at	0.547074	1.6e-26	3.4e-22	AT1G79920	heat shock protein hsp70, putative	
251683_at0.546546	1.8e-26	3.9e-22	AT3G57120	protein kinase family		
249730_at0.546272	1.9e-26	4.2e-22	AT5G624430	calcium-dependent protein kinase, putative (CDPK)		
247811_at0.545759	2.2e-26	4.8e-22	AT5G58430	leucine zipper-containing protein		
248698_at0.545115	2.6e-26	5.6e-22	AT5G48380	leucine rich repeat protein family		
264436_at0.544316	3.1e-26	6.8e-22	AT1G10370	glutathione transferase, putative		
254432_at0.543187	4.1e-26	9.0e-22	AT4G20830	FAD-linked oxidoreductase family		
252206_at0.541739	5.9e-26	1.3e-21	AT3G50360	caltractin (centrin)		
266964_at0.539990	9.1e-26	2.0e-21	AT2G39480	ABC transporter family protein		

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

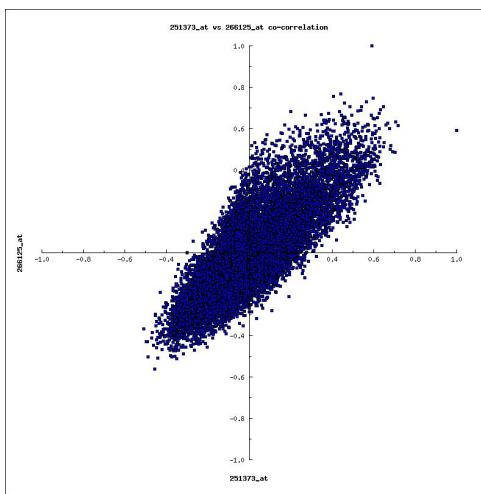


Fig. 1: Co-correlation scatter plot of GATA-2 against GATA-4.

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.