Tuberculosis (2004) **84**, 275–281

Tuberculosis

www.elsevierhealth.com/journals/tube

Visualization of microarray results to assist interpretation

Irene Papatheodorou[a], Marek Sergot[a], Marty Randall[a], Graham R. Stewart[b], Brian D. Robertson[b,*]

[a]*Department of Computing, Imperial College London, London SW7 2AZ, UK*
[b]*Centre for Molecular Microbiology and Infection, Flowers Building, Imperial College London, London SW7 2AZ, UK*

Accepted 23 December 2003

KEYWORDS
Directed graphs;
Visualization;
Microarray data

Summary Whole genome microarrays allow assessment of the profile of genes expressed under particular experimental conditions, including external stimuli such as pH or temperature, and internal changes brought about by deleting or over-expressing a gene. Such experiments produce large data sets, for which sophisticated analysis software is available. What is lacking are tools for analysing data sets from different experiments, in order to test and generate hypotheses about the links between regulatory networks. We describe here a method for presenting results from different experiments as a directed graph constructed using an automated graph drawing program *xneato*, enhanced by a logic program designed to cluster data and aid in the generation of hypotheses about possible gene interactions. A web-based front-end to the system has been constructed to explore and manipulate the graphical displays produced. Results of microarray experiments on *Mycobacterium tuberculosis* were used to develop and evaluate the visualization tool and initiate the development of an inference system for gene interactions based on such data. The GeneGraph project can be accessed at: zebrafish.doc.ic.ac.uk

Introduction

The regulation of gene expression patterns in response to internal and external stimuli is complex. Specific genes are switched on and off in a well-concerted manner so that the cell adapts to new conditions, but the pathways are often interconnected in ways which are difficult to investigate experimentally. Traditional methods for monitoring gene expression typically measure one gene per experiment and as a result are time consuming and provide limited information. More recently, the success of genome sequencing projects and the emergence of microarray technology allows the response of all the genes in the genome to be analysed simultaneously.

Microarray experiments can be designed to test a large variety of hypotheses. A typical experiment involves exposure of the cell to an environmental condition in vitro such as heat-shock or low oxygen; another form involves deletion or over-expression of a gene or combination of genes. These experiments aim at identifying genes whose expression is affected by the test conditions, with the further goal of understanding the complexity of gene interactions and unravelling the pathways that

*Corresponding author. Tel.: +44-20-7594-3198; fax: +44-20-7594-3095.
E-mail address: b.robertson@imperial.ac.uk (B.D. Robertson).

1472-9792/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tube.2003.12.012

define the cell's responses to various stimuli. With linear gene-by-gene methods the amount of data generated could be analysed by manual methods. For microarray data sets, there is a need to develop computational methods to visualize and interpret the large volumes of expression data produced.

We describe a method for presenting and visualizing the results of microarray experiments in the form of a directed graph (that is, a set of nodes interconnected with arrows). The graph is constructed by an automated graph-drawing program, a modified version of the *neato* program for producing directed graphs provided as part of the *Graphviz* package, and enhanced by a logic program designed to cluster data and make hypotheses about possible gene interactions. The web-based front-end enables the user to explore and manipulate the graphical displays produced. Our aim in the longer term is to develop inference methods for producing hypotheses about possible gene interactions as more and more data sets are accumulated; in this note we describe the visualization tool developed so far. The web interface and further details about the code are available at zebrafish.doc.ic.ac.uk.

Methods and results

Microarray data

The datasets for the microarray experiments used in the development of the tools were obtained from two sources. One source are microarray experiments performed in-house,[1] the other is the supplementary data provided by the Schoolnik lab on their website[2] to published papers.[3–6] Both sets of microarray experiments involve deletion or overexpression of genes and exposure of the wild-type and mutant strains to defined environmental conditions. The data are in the form of fold change in expression using the number of standard deviations from the control mean as a significance cut-off to determine up- or down-regulation.

All these data sets were stored in standard MySQL[7] and PostgreSQL[8] relational databases along with information from the Sanger Centre[9] about the sequenced strain *M. tuberculosis* H37Rv[10] including the predicted gene product, gene length, accession numbers in other resources such as GenBank, and a standard classification of known function.

Graph drawing tool

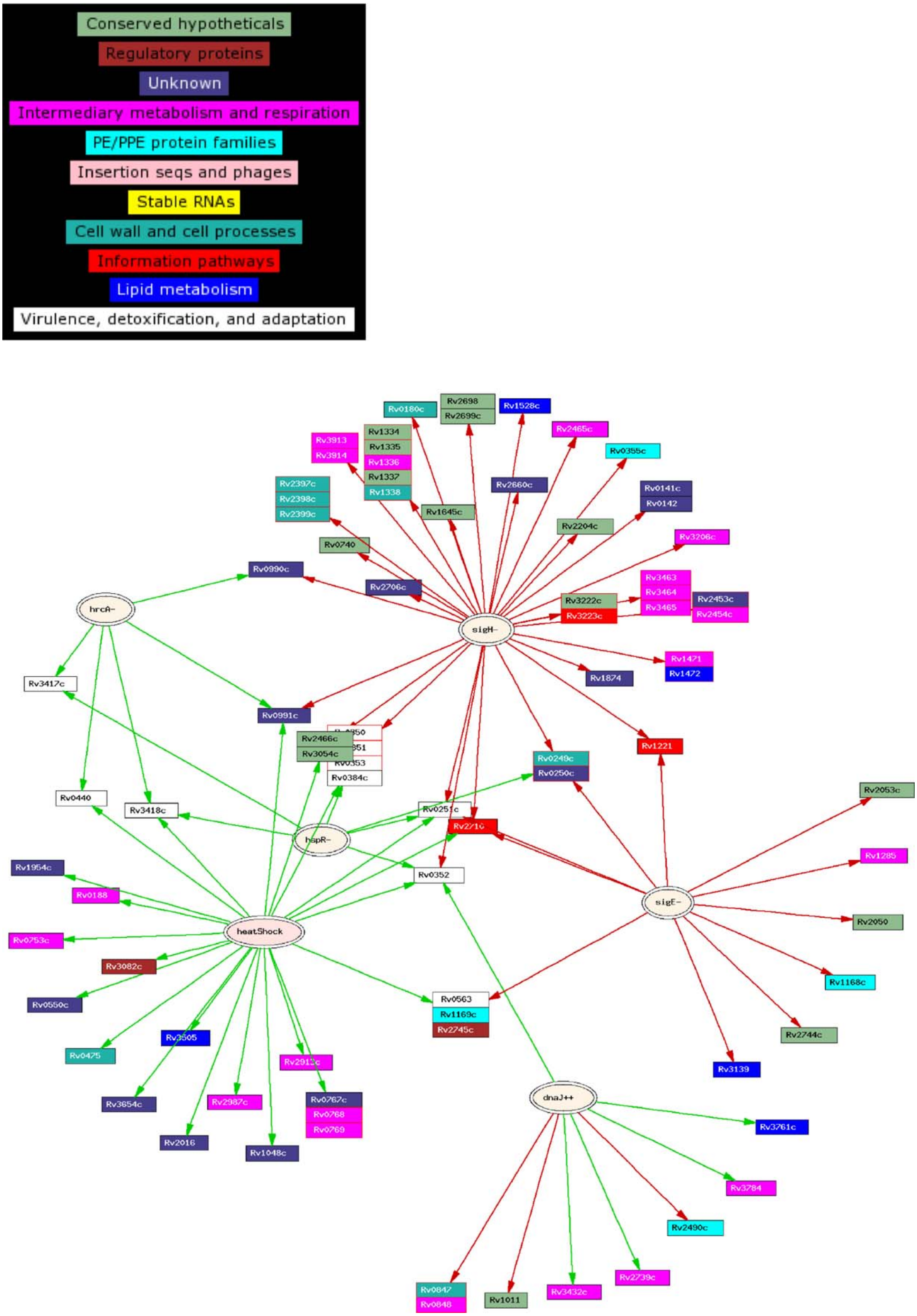
Experimental results from selected combinations of microarray experiments are presented in the form of a graph: one set of nodes represent the experiments, another set represent the genes whose expression is affected, and arrows represent the activation/repression relations between the experiment nodes and the gene-nodes. We thereby provide a graphical way of comparing different experimental results, as illustrated in Fig. 1.

The graph drawing tool we employ is *xneato*, which is a modified version of the *neato* program for the automated construction of directed graphs provided as part of the open source *Graphviz* package from AT&T Laboratories.[11] In its current version, *neato* does not provide mechanisms for drawing clusters and sub-clusters of nodes; our modified version, *xneato*, uses *neato* as the underlying graph drawing engine and extends it to produce clusters and sub-clusters of nodes. The *xneato* package is integrated with the database, where we store the experimental results and with a web-based front end providing html hyper-links to summary descriptions of experiments and gene nodes and to further information on genes and experiments available at remote sources.

Clustering

To assist in the detection of co-regulated genes, a logic program written in Prolog, is used to compare the results of different combinations of microarray

Figure 1 This graph displays data obtained from three types of experiments. (1) The *dnaJ*++ (doi:10.1016/j.tube.2003.12.009) data is the gene expression profile from a mutant over-expressing DnaJ compared to wild type. (2) The *sigH*-[6] (Rv3223c), *sigE*-[4] (Rv1221), *hspR*-[1] (Rv0353) and *hrcA*-[1] (Rv2374c) data are the gene expression profiles from strains carrying null mutations in those genes compared to wild type. (3) The *heatshock* data are the genes upregulated in response to heat-shock compared to wild type. The green arrows link the experiment node to the genes whose expression levels appear increased in the experiment, whereas the red arrows link to the genes whose expression levels appear decreased in the experiment. The colours of the gene nodes represent a standard classification by known functions. The stacks of nodes group together genes that are adjacent in the chromosome or genes that are affected in similar ways by the three experiments. When adjacent genes are also on the same strand, there is the likelihood that they belong to the same operon, as indicated by red boxes.


Conserved hypotheticals
Regulatory proteins
Unknown
Intermediary metabolism and respiration
PE/PPE protein families
Insertion seqs and phages
Stable RNAs
Cell wall and cell processes
Information pathways
Lipid metabolism
Virulence, detoxification, and adaptation
Rv0180c
Rv2698
Rv2699c
Rv1528c
Rv2465c
Rv3913
Rv3914
Rv1334
Rv1335
Rv1336
Rv1337
Rv1338
Rv0355c
Rv2660c
Rv0141c
Rv0142
Rv2397c
Rv2398c
Rv2399c
Rv1645c
Rv2204c
Rv3206c
Rv0740
Rv0990c
Rv2706c
Rv3463
Rv3464
Rv3465
Rv3222c
Rv3223c
Rv2453c
Rv2454c
hrcA-
sigH-
Rv1471
Rv1472
Rv1874
Rv3417c
Rv0991c
Rv2466c
Rv3054c
Rv0353
Rv0384c
Rv1221
Rv0249c
Rv0250c
Rv0440
Rv3418c
Rv0251c
Rv2710
hspR-
Rv2053c
Rv1285
Rv1954c
Rv0352
Rv0188
sigE-
heatShock
Rv0753c
Rv2050
Rv3082c
Rv0563
Rv1169c
Rv2745c
Rv1168c
Rv0550c
Rv3505
Rv2744c
Rv0475
Rv2912c
Rv3139
Rv3654c
Rv2987c
Rv0767c
Rv0768
Rv0769
dnaJ++
Rv3761c
Rv2016
Rv1048c
Rv3784
Rv2490c
Rv2739c
Rv0847
Rv0848
Rv1011
Rv3432c

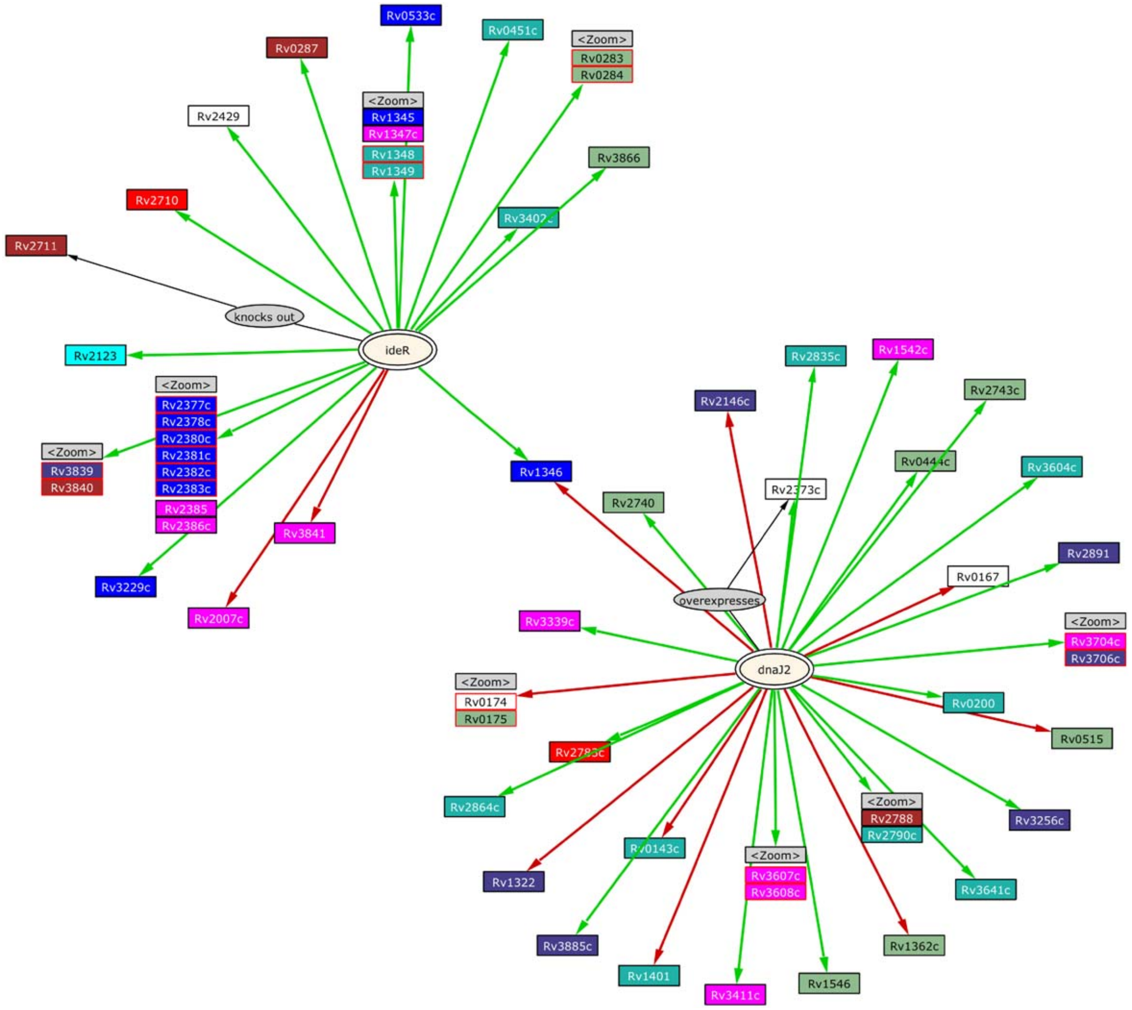


Figure 2 This graph displays the experiments that involve deletion of the gene Rv2711 (*ideR*)[5] and over-expression of Rv2373c (*dnaJ2*) (doi:10.1016/j.tube.2003.12.009) respectively. Zoom boxes refer to the online version and allow stacks to be expanded.

experiments and to cluster the genes. There are three types of clusters:

1. Genes affected in the same way by all experiments in the comparison set.
2. Genes affected by one experiment in the comparison set but adjacent along the chromosome.
3. Hypothetical operons, genes adjacent along the same DNA strand of the chromosome and affected in the same way by all experiments in the comparison set.

The clusters are shown in Fig. 1 as stacks of nodes on the graph. Hypothetical operons are depicted as stacks with red borders.

The first two types of cluster are provided mainly for presentational reasons, to reduce clutter on the graph. Directed graphs can become very crowded and difficult to interpret (and draw) with increasing number of nodes. By clustering the gene nodes we reduce the number of arrows, avoid node overlaps, and produce a much clearer graph that is easier to interpret.

The third cluster type, hypothetical operons, attempts to identify groups of spatially linked genes regulated together which may function as operons. The hypothetical operons produced as clusters of type 3 are treated as atomic units, in that they can appear in stacks of the other two cluster types. The more microarray experiments included in a comparison set, the greater the accuracy of operon prediction, since there is more evidence that those genes are expressed together.

The clustering program obtains data for the comparison set of experiments from the database, constructs the three types of clusters, and outputs information about the graph's nodes, arrows, and clusters for processing by *xneato*. *xneato* uses

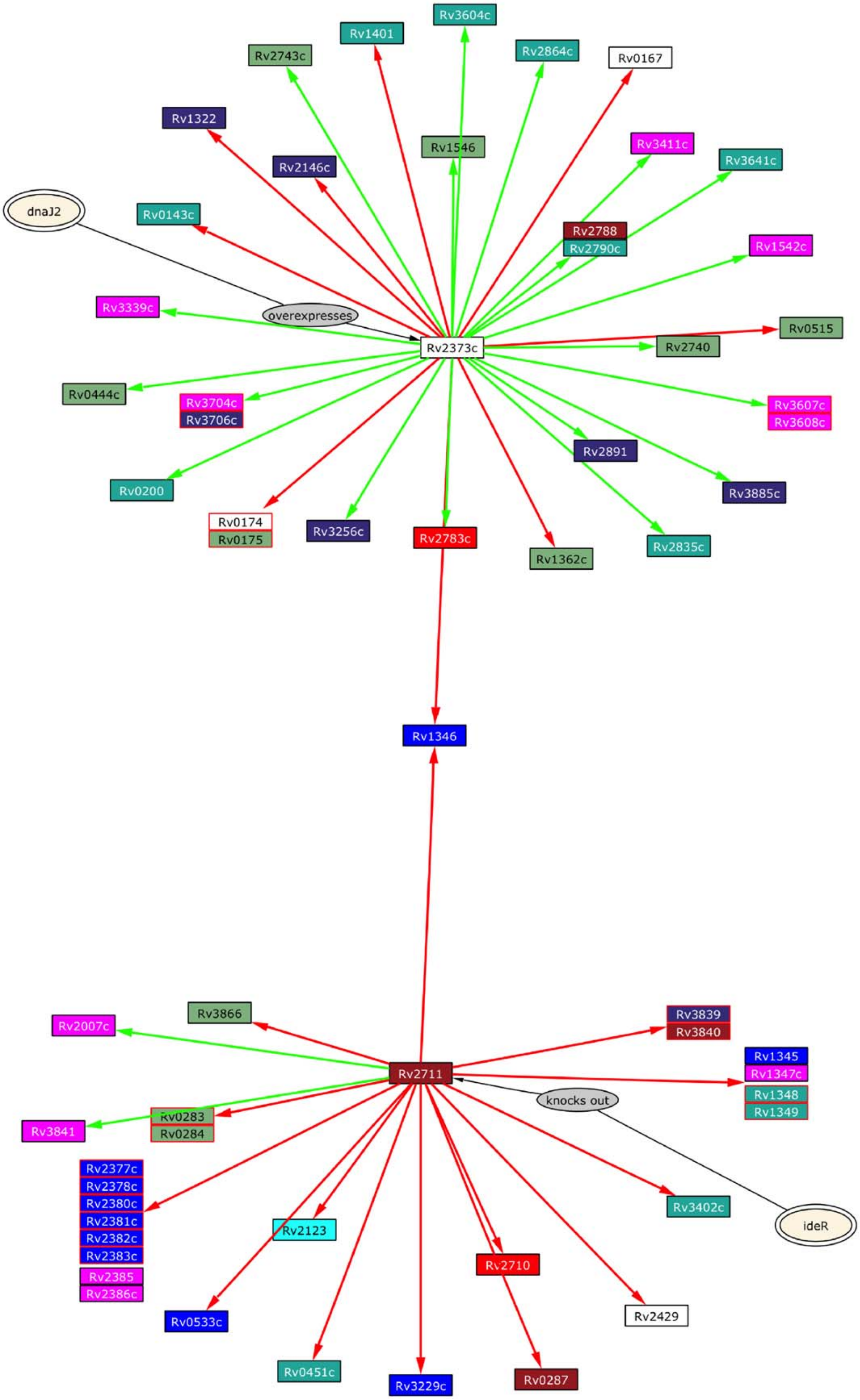


Figure 3 This graph displays the relations derived from applying the different hypotheses in the experiments in Fig. 1. The central nodes are now the genes deleted or over-expressed in the experiment.

neato to lay out and display the graph, including hyper-links that can be processed by a browser in a web-based interface. The GeneGraph web interface may be viewed at zebrafish.doc.ic.ac.uk.

Inference of relationships between genes and environmental factors

The clustering program has been further extended with the addition of a "hypothesis" option. In addition to viewing selected combinations of experimental data, the user can choose to view a graph depicting a classification of genes according to the environmental factors that affect them or by possible interactions with other genes.

This option aims to assist in the evaluation of experimental hypotheses and the design of new microarray experiments. For example, for the genes that appear up-regulated in an experiment where a certain *geneX* was mutated, we may infer that the mutated *geneX* (directly or indirectly) inhibits the transcription of those genes. This inference is obtained by formulating, in Prolog, simple inference rules that may be used to interpret the data. We have implemented inference rules that characterise defined mutants, test in vitro conditions, or test up-regulation of genes. We call these rules "First Inference Processes" since they are the initial step to the inference of interactions between gene products from microarray results.

The inference results can be viewed as a directed graph in the same way as experimental data. In these graphs, however, the central nodes are not experimental nodes but nodes representing the genes that have been mutated or over-expressed, or nodes representing the environmental factors, such as heat or oxygen, that had been altered in experiments. An example of an "experimental data" graph is shown in Fig. 2 with a "hypotheses" graph in Fig. 3.

Finally, there are cases where experiments are designed to test two different conditions at the same time, as for example where a mutant strain is exposed to low oxygen. In such experiments it is less clear which condition is responsible for differences in gene expression. Inferences about hypothetical mechanisms can be made by comparing these experiments against simpler microarray experiments varying one factor at a time, though the range of possibilities is quite wide. In this case comparison with the wild type can be used to identify the set of genes which are no longer regulated when a particular gene is deleted. Formulation of inference rules for interpretation of combinations of experiments is a main component of our current investigations.

Discussion

Our software constructs directed graphs to provide a pictorial depiction of microarray experiments, their components, and how these experiments affect the expression of genes. The web interface enables a quick identification of each gene appearing in the graph and provides links to information stored in our database or available at public databases. Colour coding of the gene-nodes classifies genes to eight main functional groups that are available at the Sanger Centre. The clustering program can successfully identify potential operons and formulate a set of initial hypotheses. Integration of the clustering program and *xneato* to the web interface enables users to select various combinations of experiments they wish to view.

Future work will focus on the implementation of inference rules dealing with the interpretation of combinations of experiments. We also aim to integrate information that is already available on known gene interactions in *M. tuberculosis*. We are in the process of obtaining data from more microarray experiments on this organism, and are planning to test the tool on a larger variety of microarray data from different organisms. Raw data are in general not readily available for microarray experiments which precludes the reanalysis of data sets using the same statistical tests. Another issue is that some knocked-out genes still give positive signals on hybridization—this may be a consequence of gene remnants hybridizing to the array, but requires details of how mutants were constructed for evaluation. It is hoped that the availability of MIAME compliant microarray data[12,13] will solve these problems.

Related techniques for the elucidation of pathways from different types of experimental data have been published.[14,15] Although scalable, the methods described are not suitable for microarray data sets. At this stage of development the tools described allow the visualization of microarray data, and how data from different experiments overlap. For example, Fig. 1 illustrates that only a subset of genes up-regulated in the heat shock response are controlled by the classic heat shock regulators, demonstrating that many of the genes are controlled by as yet unknown mechanisms. In the longer term our aim is to develop a larger set of inference rules that will be used (a) to confirm that proposed pathways do not conflict with experimental data and prior knowledge, and (b) to provide a pathway finding facility that will propose new pathways to explain experimental data and suggest new discriminating microarray experiments.

Acknowledgements

This work was funded by the Wellcome Trust (BDR and GRS).

References

1. Stewart GR, Wernisch L, Stabler R, Mangan JA, Hinds J, Laing KG, Young DB, Butcher PD. Dissection of the heat-shock response in *Mycobacterium tuberculosis* using mutants and microarrays. *Microbiology* 2002;**148**:3129–38.
2. http://schoolniklab.stanford.edu/projects/tb.html.
3. Sherman DR, Voskuil M, Schnappinger D, Liao R, Harrell MI, Schoolnik GK. Regulation of the *Mycobacterium tuberculosis* hypoxic response gene encoding alpha-crystallin. *Proc Natl Acad Sci USA* 2001;**98**:7534–9.
4. Manganelli R, Voskuil MI, Schoolnik GK, Smith I. The *Mycobacterium tuberculosis* ECF sigma factor sigmaE: role in global gene expression and survival in macrophages. *Mol Microbiol* 2001;**41**:423–37.
5. Rodriguez GM, Voskuil MI, Gold B, Schoolnik GK, Smith I. *ideR*, An essential gene in *Mycobacterium tuberculosis*: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response. *Infect Immun* 2002;**70**:3371–81.
6. Manganelli R, Voskuil MI, Schoolnik GK, Dubnau E, Gomez M, Smith I. Role of the extracytoplasmic-function sigma factor sigma(H) in *Mycobacterium tuberculosis* global gene expression. *Mol Microbiol* 2002;**45**:365–74.
7. http://www.mysql.org.
8. http://www.postgresql.org.
9. http://www.sanger.ac.uk/Projects/M_tuberculosis/.
10. Cole ST, Brosch R, Parkhill J, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 1998;**393**:537–44.
11. http://www.research.att.com/sw/tools/graphviz.
12. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball C, Causton H, Gaasterland T, Glenisson P, Holstege F, Kim I, Markowitz V, Matese J, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* 2001;**29**: 365–71.
13. Brazma A, Sarkans U, Robinson A, Vilo J, Vingron M, Hoheisel J, Fellenberg K. Microarray data representation, annotation and storage. *Adv Biochem Eng Biotechnol* 2002;**77**: 113–39.
14. Zupan B, Demsar J, Bratko I, Juvan P, Halter JA, Kuspa A, Shaulsky G. GenePath: a system for automated construction of genetic networks from mutant data. *Bioinformatics* 2003;**19**:383–9.
15. Zupan B, Bratko I, Demsar J, Juvan P, Curk T, Bortšnik U, Beck R, Halter J, Kuspa A, Shaulsky G. GenePath: a system for inference of genetic networks and proposal of genetic experiments. *Artif Intell Med* 2003;**29**:107–30.