CASE STUDY

Visualising gene expression in its metabolic context

Detlef Wolf, Christopher P. Gray and Antoine de Saizieu

INTRODUCTION

Transcription profiling and proteomics are technologies that allow the measurement of expression levels for many genes in parallel (in our case of ~2,000 and ~500 genes, respectively). Biologists require tools to assist in the interpretation of the large sets of data obtained. Popular methods include scatter plots of pairs of experiments (Affymetrix1) and clustering (hierarchical, Kohonen-map) with subsequent visualisation.2,3 Common to these methods is the establishment of distance measures between pairs of genes. The distance measure is used to recognise clusters and then to interpret the clusters based on gene annotation.

A system for mapping gene expression data onto pathways is also described.4 To our knowledge this system seems to have advantages in showing overview pathway diagrams while the system described below allows integrated visualisation of data quality and series of experiments.

In the approach described here we accept the distances and contexts given by known metabolic pathways and map the expression levels onto these pathways. In this manner we enhance our current understanding of both expression data and the pathways. This paper has two parts, the first of which comprises examples of biological applications, and the second describes the underlying software system.

RESULTS

Haemophilus influenzae, Affymetrix chips and KEGG pathways

Haemophilus influenzae cells, grown in full medium, were first washed in starvation media to remove the brain-heart-infusion (BHI) media (which contains a high concentration of RNA) and were then incubated in starvation medium for various times. RNA was extracted and cDNA synthesized incorporating a biotin label. The cDNA was then used as a target for Affymetrix...
Figure 1: Affymetrix expression data of H. influenzae projected onto a KEGG pathway diagram. The colour of the left half of each coloured box is determined by the data from the first replicate of the experiment, the data for the right half comes from the second replicate of the experiment.
Chips containing probes for all known *H. influenzae* genes. In the subsequent text these samples will be referred to by the incubation time.

The procedure was repeated under the same conditions except that the cells were kept at 0°C during the washing procedure. RNA extracted immediately after the washing of the cells under these conditions is referred to as 'cold' as opposed to 'warm' for RNA extracted after washing the cells at room temperature.

The whole experiment was repeated, resulting in two pairs of chip data. The ratios warm/cold of the normalised Affymetrix signal values were projected onto KEGG (Kyoto Encyclopedia of Genes and Genomes) metabolic pathway maps\(^5\) using the software described below. Enzyme boxes that could be associated with one or more *H. influenzae* genes are coloured, otherwise a box remains white. The left half of each enzyme box was coloured using data from the first replicate of the experiment, the right half using data from the second experiment. Two small tick marks close to the upper left corner of an enzyme box are drawn if more than one gene has been annotated with this EC number; the procedure for selecting one of the possible expression values for colouring this case is described below. One small tick mark close to the lower left corner indicates that the Affymetrix signal value (average difference of probe pairs) for at least one of the two replicates of the experiment was below 300 (Figures 1 and 2).

In purine metabolism, a significant up-regulation of most of the genes encoding the enzymes involved in the synthesis of Inosine-5'-Phosphate (IMP) from either l-glutamine or l-glutamate is observed in both experiments. This result is visually appealing because the up-regulated genes are contiguous in the diagram (not on the genome) whereas most of the remaining genes do not significantly change, the exception being those encoding the subunits of DNA polymerase (holA, holB, holC and holD).

**Streptococcus pneumoniae**, Affymetrix chips and KEGG pathways

*Streptococcus pneumoniae* cells were grown in rich medium in a flask. Four samples were collected during the growth phase at optical densities 0.1, 0.3, 0.5 and 1.0. In parallel, cells were grown to high density in a sorbarod filter (biofilm simulation). Projecting the Affymetrix measurements of these five experiments onto the KEGG pathway diagram of tRNA synthetases (Figure 3) suggests that the translation machinery is generally down-regulated in stationary...
phase when compared to exponentially growing bacteria (eg EC 6.1.1.5, isoleucine-tRNA synthetase). The translation machinery for biofilm-like conditions and stationary phase was found to be regulated in a similar manner. This type of visualisation allows the rapid identification of values for which the data are questionable owing to low signal (eg EC 6.1.1.19) and for values that contradict hypotheses (eg EC 6.1.1.16 has a strong signal indicating down-regulation under biofilm-like conditions).

**Haemophilus influenzae**, 2D-gels and custom pathways

Custom pathway diagrams consisting of metabolite names and vectors symbolising reactions for which the corresponding spot(s) on 2D gels are known were drawn manually. Two cultures of *Haemophilus influenzae* were induced with Trimethoprim (Tmp), one for 10 minutes and the other for 30 minutes, before labelling for 3 minutes with ^35^S. Induction ratios, compared to the non-induced parallel cultures, were calculated for the intensities of the spots representing the respective enzymes after 2D-PAGE (polyacrylamide gel electrophoresis) analysis and used to automatically colour the arrows in custom pathway diagrams (Figure 4). Induction of proteins that are involved in the biosynthesis of methionine and require tetrahydrofolate (THF) was the major observation. Both diagrams show unconfirmed repression of the enzyme catalysing the interconversion of oxaloacetate and phosphoenol pyruvate. As both diagrams were consistent in this respect the confidence in this result can be relatively high. However as both ratios were based on the same control experiment the reproducibility of this spot in the control gel(s) would have to be examined.

Displaying one diagram at a time at a fixed screen location so that switching between diagrams alters the colours while keeping the arrows in place provides a simple method to visualise changes across a series of
Visualising gene expression in its metabolic context

**Figure 4:** Proteomics expression data for *H. influenzae* after 10 minutes of TMP induction. The size of the coloured squares on the arrows increases with the uncertainty of the value measured. Small black squares indicate bad reproducibility or unconfirmed results.

**Diagrams** (in this example ordered by induction time).

**Software**

**KEGG/Affymetrix**

As Affymetrix chips are organised according to gene identifiers while KEGG metabolic pathway diagrams are organised by EC numbers in given positions, the key element in linking expression data with KEGG diagrams is the many-to-many mapping between genes and EC numbers.

One gene can encode proteins possessing multiple enzymatic functions (e.g. *purH* (HI0887) from *H. influenzae* potentially catalyses reactions of types 2.1.2.3 and 3.5.4.10) and one enzymatic function can require the protein products of several genes (e.g. 2.7.7.7 (DNA-directed DNA polymerase) is associated with nine *H. influenzae* genes).

The mapping between reactions and EC numbers is also many-to-many, represented in KEGG by several EC numbers drawn next to a reaction arrow. One EC number can also occur in several places on one or more KEGG pathway diagrams.

While the many-to-many nature of the gene/EC number relationship is reasonably well treated in the colouring software, the many-to-many nature of the reaction/EC number relationship is currently beyond the ‘biochemical resolution’ of KEGG and also not yet addressed by our software.

In the case of *H. influenzae* the gene/EC number mapping provided by KEGG was used whereas for the *S. pneumoniae* sequence a mapping was predicted based on BLAST sequence similarity search against SWISS-PROT. BLAST output was parsed and filtered based on percentage sequence identity and length. A BLAST hit was accepted if the sequence identity was better than 60 per cent for alignments of 30 or more residues in length or the identity was better than 40 per cent for lengths of 50 or more residues. Subsequently EC numbers were extracted from the SWISS-PROT sequence description lines (DE).
Affymetrix chip experiment results are accessed at the level of the so-called ‘chip file’ which contains, for each gene, a set of numbers one of which is designated ‘average difference intensity’. This number correlates with the amount of mRNA present in the sample originating from a particular gene. We use the Roche Affymetrix Chip Experiment Analysis system (RACE-A, by Clemens Broger and Martin Neeb) to compute averages across replicates of chip experiments and ratios between different experimental conditions (Figure 5).

The program keggcolor reads pathway diagrams from KEGG, the gene/EC number mapping and the expression data and uses a routine that can change all pixels in a defined rectangle that have colour A into colour B. Colour A, the white background of KEGG GIF images, is changed into a colour corresponding to the expression level according to a predetermined colour scale.

When the program keggcolor is invoked, a list of comparison names between Affymetrix chip experiments (where replicates are combined into one experiment) is passed to the program. Each comparison name has the form C_T (control experiment–underscore–test experiment). This allows the program to derive the list of individual experiment names from the list of comparisons.

If several genes map to one EC number, then

(1) genes are selected that show at least an average difference of 300 (which is believed to be reliable) in all individual experiments;

(2) among these the one gene with the highest average absolute change across all comparisons is selected for colouring the EC number box;

(3) if no highly expressed gene exists, rule (1) is dropped, and a tick mark is drawn in the lower left corner of the box.

Routinely we do not colour all KEGG maps, but only the ones with a high number of mapped enzymes. Table 1 shows the top six KEGG pathway maps for which S. pneumoniae gene/EC number assignments could be constructed.

Table 1

<table>
<thead>
<tr>
<th>KEGG pathway diagram</th>
<th>Count of different EC numbers</th>
<th>Count of different EC numbers mapped to genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phospholipase, tyrosine and tryptophan biosynthesis</td>
<td>31</td>
<td>17</td>
</tr>
<tr>
<td>Starch and sucrose metabolism</td>
<td>73</td>
<td>17</td>
</tr>
<tr>
<td>Glycolysis/gluconeogenesis</td>
<td>35</td>
<td>18</td>
</tr>
<tr>
<td>Aminoacyl-tRNA biosynthesis</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>Pyrimidine metabolism</td>
<td>59</td>
<td>24</td>
</tr>
<tr>
<td>Purine metabolism</td>
<td>91</td>
<td>30</td>
</tr>
</tbody>
</table>
KEGG diagrams are constructed from generic biochemical pathways and annotated genes from a set of species which results in the inclusion of all known enzymes (EC numbers) for any reaction. Many gene annotations have been inferred in silico based only on sequence homology. This leads to additional EC numbers assigned to reactions in KEGG diagrams and explains the high number of different EC numbers in, for example, the KEGG diagram for purine metabolism.

Figure 6 shows the software components written or extended in this context. Both KEGG coordinates and the gene/EC number mapping as well as the Affymetrix chip experiment results are stored in ORACLE, the KEGG GIF drawings are kept in files.

**Proteomics/In-house diagrams**

An in-house system is used to store the values obtained for gel spot intensities, to combine replicate gels and to compute ratios between control and test gel replicates. Custom pathway diagrams are drawn using, for example, MS-PowerPoint. The program ‘picture annotator’, containing a Java applet that is connected to an ORACLE database via a common gateway interface (CGI) program, is used to manually record the relationship between shapes in the diagram and masterspots on the consensus gel. It is possible to enter either the gene identifier or, if a gene is represented by several masterspots, a set of masterspots. A flood-fill algorithm is used to recolour the whole shape originating from the point indicated by the user (Figure 7).
In addition a small square is used to visualise the quality of a measurement: the smaller the square, the smaller the measurement error. For reasonably small errors, the square is drawn in the colour of the shape. For large errors, missing or unconfirmed measurements (eg only one spot in the control or treated gel group) a small white square is drawn.

Some enzymes are represented by several spots on the gels. The spot that changed most is selected to colour the shape. For the visualisation of the measurement quality, the worst case is assumed: the minimum quality value is used.

CONCLUSIONS

A simple method for visualising expression data on KEGG and custom pathway diagrams has been presented which allows a quick evaluation of the consistency of expression data and the degree of agreement with present pathway knowledge.

A useful extension of the system would be to dynamically combine (parts of) several pathway diagrams into one diagram (eg in the example of S. pneumoniae described above, 60 per cent of all the housekeeping genes are down-regulated in stationary phase, but these genes are dispersed over several KEGG pathway diagrams). In such a combined diagram it would be helpful if subpathways could be collapsed into single nodes. The number of coloured pathway maps increases both with the growing number of experimental conditions and pathway diagrams. An automatic filter that recognises coloured pathway diagrams with ‘interesting’ patterns would also be helpful, eg the computer could select diagrams containing clusters of more than five genes that are both co-expressed and connected via a given maximum number of pathway steps.

A useful variation of the theme described above would be the projection of expression data onto graphical representations of the genome and synchronisation of the pathway and the genome displays via the Model/View/Controller pattern.

Acknowledgement

Clemens Broger (F. Hoffmann-La Roche, Basel) contributed the sequence analysis results for S. pneumoniae.

References

1. GeneChip® is a registered trademark of Affymetrix Inc., Santa Clara, CA 95051, USA; URL: http://www.affymetrix.com/