Pathway Based Microarray Analysis, Utilising Enzyme Compounds and Cascade Events

S. Pavlidis; S. Swift; A. Payne
School of Information Systems, Computing and Mathematics, Brunel University, London, UK

Keywords
Microarray analysis, biochemical pathways, computational biology, computing methodologies

Summary
Background: Pathway based microarray analysis is an effort to integrate microarray and pathway data in a holistic analytical approach, looking for coordinated changes in the expression of sets of genes forming pathways. However, it has been observed that the results produced are often cryptic, with cases of closely related genes in a pathway showing quite variable, even opposing expression.

Objectives: We propose a methodology to identify the state of activation of individual pathways, based on our hypothesis that gene members of many pathways or modules exhibit differential expression that results from their contribution to any combination of all their constituent pathways. Therefore, the observed expression of such a gene does not necessarily imply the activation state of a given pathway where its product participates, but reflects the net expression resulting from its participation in all its constituent pathways.

Methods: Firstly, in an effort to validate the hypothesis, we split the genes into two groups; single and multi-membership. We then determined and compared the proportion of differentially expressed genes in each group, for each experiment. In addition, we estimated the cumulative binomial probability of observing as many or more expressed genes in each group, in each experiment, simply by chance. Second, we propose a hill climbing methodology, aiming to maximise the agreement of gene expression per module.

Results: We detected more frequent expression of multi-membership genes and significantly lower probabilities of observing such a high proportion of differentially expressed multi-membership genes, as the one present in the dataset. The algorithm was able to correctly identify the state of activation of the KEGG glycolysis and gluconeogenesis pathways, using a number of Saccharomyces cerevisiae datasets. We show that the result is equivalent to the best solution found following exhaustive search.

Conclusions: The proposed method takes into account the multi-membership nature of genes and our knowledge of the competitive nature of our exemplar modules, revealing the state of activity of a pathway.

1. Introduction

Biological processes involve protein encoding genes which interconnect into regulatory networks, forming pathways. Pathway based microarray analysis focuses on the physical and functional interactions between these genes [1]. It employs gene expression analysis to identify coordinated changes occurring in the expression of sets of genes acting in concert, such as those forming biochemical pathways [2]. The balanced function of these networks is crucial for the normal functioning of an organism, e.g. deregulation of signalling cascades has major involvement in pathogenesis, notably in cancer, hence, differential expression analysis, in terms of pathways, has drawn considerable interest [3].

Research in the field relies heavily on pathway databases, cataloguing genes, proteins and chemical compounds involved in pathways. The Kyoto Encyclopaedia of Genes and Genomes (KEGG) database [4] is arguably the most popular, along with MetaCyc [5], Reactome [6] and others. Tools utilising microarray and gene sets data for expression analysis include EuGene [7], GenMapp [8], Cytoscape [9] and so on. They provide useful pathway visualisations, often employing statistical methods to score gene sets according to the degree to which the experimental conditions influence their activity. However, genes forming a pathway often show contradicting expression, with some being up-regulated while others down-regulated, while at the same time the number of expressed genes in a path is often low [1], making it difficult to infer the state of activation of a pathway. Presumably, this is due to a number of reasons, including the regulation of protein function at the level of translation and post-translational modifications [10], besides transcription, and the...
interconnection of pathways in context-specific manner, where the resulting networks target and change only a few genes per pathway [1]. For example, in a dataset of 5000 random microarray experiments obtained from Gene Expression Omnibus (GEO) [11], we observed that around 20% show high level of contradictions with more than 50 genes showing opposing expression while in the same pathway.

We argue that there is an additional cause that explains this observation, related to the fact that a significant proportion of KEGG genes are members of a number of pathways. For example, in *Escherichia coli* and *Saccharomyces cerevisiae* about 35% of genes placed in KEGG pathways have multiple memberships. Figure 1 shows the number of *Saccharomyces cerevisiae* genes of different membership degrees according to KEGG.

Our hypothesis is that the net expression of multi-membership genes may result from their contribution to any combination of their constituent pathways. Hence, it may mask the true state of a pathway when examined in isolation. Thus, we proposed a heuristic search methodology to acquire an educated guess regarding the state of individual pathways, by maximising the agreement in terms of expression of genes per pathway [12, 13]. In this paper we seek to further develop and validate our hypothesis as well as improve and refine this methodology, using the Glycolysis/Gluconeogenesis KEGG pathway as a model and proof of concept. The notion of enzyme positions, the KEGG Glycolysis and gluconeogenesis modules, consisting of a cluster of functionally related genes, which are candidates of having similar expression, and the fact that a number of genes may encode the same enzyme is taken into account, as discussed in detail in the following section.

2. Rationale

Our rationale can be exemplified on Figure 2, showing a diagram of the *Saccharomyces cerevisiae* KEGG glycolysis gluconeogenesis pathway, consisting of the respective KEGG glycolysis and gluconeogenesis modules. Rectangles represent positions where enzymes catalyse a step of the pathway. The numbers represent enzyme commission (EC) numbers, a numerical classification of enzymes [14]. The substrates and products of reactions are represented by circles. As the arrows indicate, some enzymes can catalyse a reaction in both directions, while others only towards one product, such as the case of EC 4.1.2.13 and EC 2.7.1.11 respectively. Additionally, certain enzymes can catalyse more than one reaction, like EC 5.3.1.9.
We attempt to maximise the agreement of expression of EC positions per module. The membership of genes to EC positions and modules is based on KEGG and represented as discussed in the next section. In addition, in this work the algorithm assigns or allocates genes to catalytic positions (EC), for a particular module. Naturally, a gene can only be allocated to a position if it is a member of that position according to KEGG. Within this context, assigning a gene to an EC position, in a module, suggests that the state of differential expression of that gene is related to the state of that position, while removal implies no relationship. In this way we aim to elucidate which particular reaction, in which module is related to the activation or repression of a particular gene. The proportion of EC positions in a module carrying assigned genes constitutes that module's coverage.

Given that this work is currently confined to the Glycolysis/Gluconeogenesis KEGG pathway the membership of some genes in other pathways and modules is discarded. Hence, one might argue that the observed expression of such genes, in the discussed experiments, may be due to the biological system regulating their expression to satisfy the function of modules other than glycolysis and gluconeogenesis, which is not considered in the current analysis. To account for this omission this work is mainly centred on experiments for which the accompanying literature provides clear analysis of the experimental conditions and the state of the glycolysis and gluconeogenesis modules. For example we consider conditions such as addition of glucose, which have a strong effect on these biochemical processes and the accompanying publications confirm activation or repression of each one. Additionally, we used data where statistical analysis by both Eu.Gene and the method described by [18] and implemented in [13], reveal significant enrichment of the pathway in differentially expressed genes (p < 0.01). Given that we work on some time series data, where, particularly at the initial stages, the cells have not yet responded to the environmental perturbations, we comment on the instances where the above restrictions do not apply.

Due to the size of the search space, an exhaustive search is not an adequate approach for large networks, e.g. the entire metabolic network of an organism. Thus, we have opted for a Hill Climbing algorithm [19] that changes the possible multi-membership gene allocation to EC positions and modules. To examine the performance of the search we also carried out comparative analysis of the allocation produced by the hill climb to an exhaustive search.

3.1 Algorithm

Let $Y$ be a list of integers, $\{1, ..., N_1\}$, representing unique genes, let $Z = \{1, ..., N_2\}$ be the EC position identifier (that is, an integer, e.g. 4.2.1.0 might be ID: 7) where EC position identifiers that appear more than once have their own ID, and $D$ a list of module identifiers $\{1, ..., N_m\}$.

We define a list of 5-tuples where each 5-tuple represents gene $g \in Y$, enzyme (EC) $e_i \in Z$ encoded by the gene, the module.
Methods Inf Med 4/2012 © Schattauer 2012
326 S. Pavlidis et al.: Pathway Based Microarray Analysis, Utilising Enzyme Compounds and Cascade Events

Table 2 Example of a row in the list of 5-tuples, where a down-regulated gene with ID 6 is assigned to EC position with ID 10 in module 1

<table>
<thead>
<tr>
<th>gi</th>
<th>ej</th>
<th>mi</th>
<th>xi</th>
<th>li</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>10</td>
<td>1</td>
<td>–1</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

$X_i \in D$ where they participate, the expression $x_i$ of the gene and the state of allocation $l_i$ of the gene to the particular enzyme and module, as shown in Equation 1. Hence, the list of 5-tuples is based on KEGG, revealing the membership of genes to modules and EC positions, which does not change. A gene can only be allocated to a position and module if it is a member of that position and module.

$b_i = (g_i, e_i, m_i, x_i, l_i)$                      (1)

$x_i$ has a value of $+1$, $–1$ or $0$ if gene $i$ is up-, down-regulated or stable respectively, based on a threshold parameter $t$ (Eq. 2).

$x_i = \begin{cases} +1, & \text{if } G(i) > t \\ -1, & \text{if } G(i) < -t \\ 0, & \text{otherwise} \end{cases}$ (2)

$l_i = 1$ if gene $g_i$ is allocated to EC position $e_i$ in module $m_i$, and $l_i = 0$ otherwise. Thus, we work with a binary string $A$ of size equal to the length of the list of 5-tuples, representing the allocation of genes to EC positions and modules. That is, a list representing $l_i$, or the last column in our list of 5-tuples.

For example, let’s assume that gene PGK1 (say $g_6 = 6$) corresponding to EC:2.7.2.3 (say $e_1 = 10$) is down-regulated in an experiment of interest (hence $x_i = –1$) and that the gene is considered active in the glycolysis module (hence $m_1 = 1$, where 1 is the module identifier), that is, the algorithm has allocated the gene to glycolysis ($l_i = 1$). In this example, in the list of 5-tuples, the row corresponding to this gene in the corresponding EC position and module would appear as shown in Table 2.

For a particular EC position we define a scoring function $E(i)$ as shown in Equation 3, where the score is calculated using Equations 4 and 5, for up- and down-regulated genes respectively.

$E(i) = \begin{cases} 0, & |U_i| = |D_i| \\ -1, & |D_i| > |U_i| \\ +1, & |U_i| > |D_i| \end{cases}$ (3)

$U_i = \{b_j : l_j = 1 \wedge e_j = i \wedge x_j = 1\}$ (4)

$D_i = \{b_j : l_j = 1 \wedge e_j = i \wedge x_j = -1\}$ (5)

Therefore, if there are more up-regulated than down-regulated genes in a particular EC position, in a module, the position has a score of 1 and –1 otherwise, while empty positions have a score of 0. Equation 6 allows us to identify EC positions in each module, while Equation 7 assigns 0.5 to positions holding equal numbers of up- and down-regulated genes.

$K(i) = \{e_j : l_j = 1 \wedge m_j = i\}$ (6)

$H(i) = \begin{cases} 0.5, & |U_i| = |D_i| > 0 \\ 0, & \text{otherwise} \end{cases}$ (7)

Then we calculate a score for a module, based on the state of EC positions in that module, using Equation 8. Equation 9 sums positions holding more up- or down-regulated genes.

$E = \sum_{\forall e \in K(i)} E(e)$ (8)

$H = \sum_{\forall e \in K(i)} H(e)$ (9)

The more EC positions of similar expression in a module the higher the score we acquire, using Equation 10.

$M(i) = C(i)(E + Sgn(E) \times H)$ (10)

Equation 11 defines if the sum of positions of equal number of up- and down-regulated that we add in Equation 10 is positive, which is true if the module contains more up-regulated positions, or negative, when the module holds more down-regulated positions. This is biologically meaningful, as a path may switch from using the product of one gene to another, for the same position $m_i$, for the biological system to meet its needs [20]. Thus cases of equally up- and down-regulated EC positions are not in direct contradiction with our rationale.

$Sgn(X) = \begin{cases} +1, & X < 0 \\ -1, & X > 0 \\ 0, & X = 0 \end{cases}$ (11)

$C(i)$ is the module coverage, acquired by Equation 12, equal to the number of positions with at least one allocated gene divided by all enzymes/positions in the module. Naturally, the better the allocation fits the module the higher the value of $C$. Interestingly, this can be extended to routes, going through the module, when there are alternative options. We can examine how well an allocation fits each possible route.

Fig. 3 Average gene expression, for different membership degree thresholds
route. In the case of the glycolysis and gluconeogenesis modules, a plausible route starts with α-D-glucose while another with β-D-glucose. More complicated diversions are present elsewhere in the metabolic network.

\[ C(i) = \frac{\left| \{ e_j : l_j = 1 \land m_j = i \} \right|}{\left| \{ e_j : m = i \} \right|} \]  

(12)

For example let’s assume that a module consisting of 5 positions holds three positions that have been allocated more up-regulated genes and one with the same number of positively and negatively expressed genes. Using the equations above we acquire a score of 3.5, summing ones for the three positions with more up-regulated genes and 0.5 for the later one, which is multiplied by the module coverage, equal to 4/5. Consequentially, using equation (10) we would acquire a value equal to 3.5 \times (4/5) = 2.8.

We employ the fitness function (13) to search for the best allocation of genes to EC positions and modules.

\[ F(i) = \sum_{j=1}^{n} |M(i)| \]  

(13)

Algorithm 1 (Fig. 4) presents the pseudo-code for the proposed Hill Climbing search. Single-membership genes are allocated first. In this way they serve as guides for the subsequent allocation of multi-membership genes. Importantly, we disallow allocation of expressed genes to both modules given their opposing nature.

4. Hypothesis Validation

We performed comparative analysis of the frequency of expression of single- and multi-membership genes in a dataset compiled of 5000 random experiments obtained from GEO. In particular, we examined the number of experiments where each unique single- and multi-membership gene appears differentially expressed and the mean value corresponding to the totality of KEGG genes in each group. For example, if one gene is expressed in three out of six experiments and another in four out of six experiments, the average expression of these genes is \((3/6 + 4/6)/2 = 0.583\), for the group of both. Following this approach we observed that single membership genes show an average expression of 13.6% while a multi-membership genes 16.5%, in all 5000 experiments. A two sample t-test revealed that the proportions of expressed single- and multi- membership genes per experiment differ significantly, with p-value of \(6.6 \times 10^{-9}\). Importantly, the difference increases further when we consider genes with higher degree of membership, meaning genes constituting members of at least three pathways, genes present in at least four biochemical pathways and so on, as exemplified on Fig. 3. This makes sense given that the expression of multi-membership genes is regulated to contribute to a number of pathways. Hence, they are more likely to be employed by the biological system at any given instance.

Additionally, we examined the cumulative binomial probability of observing as many or more expressed genes in the group of single- and multi-membership genes, simply by chance. Here the probability of each expressed gene to fall into one of the groups is equal to the number of genes in the group divided by the overall number of gene members of pathways. In each individual experiment the overall number of trials corresponds to the number of expressed genes, while success to the number of expressed genes in the group.

A two-sample t-test showed that the probability values for the two groups differ significantly. The mean probability for single-membership genes is equal to 0.744 as opposed to 0.356 for multi-membership genes. The obtained result, shown in Fig. 5, with experiments not ordered in any particular fashion, suggests that the latter genes are more likely to show differential expression. Again, this is a sensible result given that they encode proteins of multiple functions. Thus, their expression results from their contribution to all their constituent pathways.
5.1 Heuristic Search Results

We analysed microarray data from a number of experiments on Saccharomyces cerevisiae. The algorithm correctly identified the state of the KEGG glycolysis/gluconeogenesis pathway, pinpointing which of the two modules is activated or repressed.

In [7] the authors analyse the global expression of yeast during Diazix shift, where cells inoculated in glucose rich medium turn to aerobic utilisation of ethanol. They identify the glycolysis/gluconeogenesis pathway as one of the most activated pathways in the experiment, with p-value of 0.01 at time point 7. In this approach it is unclear if the activation means increased synthesis or utilisation of glucose. Our method reveals that gluconeogenesis is activated while glycolysis repressed, in agreement with the biological analysis of the data by [21]. The method adequately identifies as fittest configuration the allocation of genes to the glycolysis module, covering eight out of 13 EC positions, so that seven appear deactivated and one equally affected. Up-regulated genes are assigned to gluconeogenesis, covering three out of eight EC positions in that module. The low coverage in this case is not unexpected and in agreement with our rationale. With cells rerouting the flow of metabolites, genes members of both the glycolysis and gluconeogenesis modules, need to be expressed in a way that balances two competing trends, repression in the first case and activation in the later. Without processing of the data, gluconeogenesis contains five down-regulated, as compared to three up-regulated genes, which would incorrectly suggest repression. Figure 6 visually portrays the result.

In the experiment with GEO ID number GSM95012 cells are subjected to a glucose pulse and microarray analysis performed on RNA extracted after 20 minutes [22]. There are 12 differentially expressed genes in the KEGG Glycolysis/Gluconeogenesis pathway. Upon processing of the data the glycolysis module appears activated with seven out of 13 positions covered, while the gluconeogenesis module repressed with four out of eight positions covered, as expected, given that cells are presented with excess glucose to cover their nutritional needs. If we were to base the analysis on the behaviour of genes only, for each module in isolation, we would observe five up-regulated and only four down-regulated genes in the gluconeogenesis pathway. Naturally, it is unclear if the module is activated or suppressed and if we were to make a sensible guess activation would be more appealing, even though this is clearly not the case here.

The results are similar for GSM94996 from the same dataset, for RNA extracted 120 minutes after the admission of a glucose pulse (2 g/l). The analysis identifies glycolysis as the activated module, with coverage of six out of 13 EC positions. In contrast, gluconeogenesis shows downward trend, given the increase of available glucose in its environment. Notably, the number of down-regulated is half the number of up-regulated genes in the gluconeogenesis group, six against three, respectively. Consequentially, examining the expression of these genes in isolation, without considering their participation in the glycolysis module, would again suggest increased synthesis of glucose. The results for 2 g/l glucose pulse between 10 and 120 minutes are exhibited on Figure 7.

In [23] the authors identify the repressive effect of nitrogen depletion on the cluster of Saccharomyces cerevisiae glycolytic genes, throughout nine consecutive time points. Our method correctly assigns differentially expressed genes to the gly-
colysis and gluconeogenesis modules, so that the former appears suppressed and the latter activated. The coverage shows a gradual increase, with evident correlation to time (0.79, p-value = 0.01) as shown on Figure 8.

Additionally, we examined the performance of the algorithm on GSM290980, which deals with the response of yeast cells to glucose deprivation [24]. In this case we expected yeast cells to switch on the gluconeogenesis process and at the same time deactivate the glycolysis module. Indeed, the method identified activation of gluconeogenesis, in agreement with biological rationale, with coverage of four out of eight EC positions. At the same time the glycolysis module appears severely repressed with coverage of all 13 EC positions, which appear down-regulated. This is expected given the lack of available glucose for degradation.

Here, there are nine down-regulated and five up-regulated gluconeogenesis genes. Looking into the module in isolation, without taking into account the participation of many of its genes in glycolysis, would lead to the assumption that glucose synthesis is suppressed. However, the methodology produces a clearer setting allowing us to detect increase of glucose synthesis and decrease of glucose degradation.

For the data discussed in this section we examined the convergence of the algorithm. Figure 9 exhibits the mean convergence for 20 separate runs of the algorithm on the Diauxic shift time point 7 data, GSM94996 and GSM290980. Naturally, the larger the search space, as defined by the number of expressed multi-membership genes and the possible allocations, the larger the number of iterations required by the algorithm to converge. Processing of GSM290980 with the most expressed genes converges after 104 iterations on average, as opposed to the time point 7 diauxic shift data, converging after 40 iterations on average.

Interestingly, the result produced by separate runs of the search on the same data where highly similar or identical. While this is encouraging, it remains to be examined how the results vary when a more complicated setting such as the entire metabolic network is used for the analysis.

### 5.2 Exhaustive Search

To examine the performance of the algorithm, we carried out comparative analysis of the allocation produced by the hill climbing technique to an exhaustive search of all possible allocation states, in terms of fitness. For each multi-membership gene we have a number of alternative possible allocations. In most cases genes are members of an EC position in two modules. Given that we only assign an expressed gene to one of the competing modules, there are two alternative allocations per gene. For a gene that is present in two positions in the same module the number grows further, since it can be assigned to any combination of these positions, excluding non-assignment. For example, a gene present in two positions in the same module we have three possible states, meaning allocation to one, the other or both. In general, the number of possible combinations is $2^n - 1$. Hence, the overall number of cases $P$ that need be con-
sidered in an exhaustive search can be estimated by multiplying the number of possible allocations for each expressed gene, according to Equation 14, where \( n \) is the number of genes and \( s \) the number of EC positions to which the gene can be assigned.

\[
P(i) = \prod_{i=1}^{n} (2^{s_i} - 1)
\]

Experiment GSM290980 is used as a case study since there are 16 expressed multi-membership genes in the Glycolysis and Gluconeogenesis modules, which is the largest set of expressed genes in the discussed experiments. Here one gene, PGI1, can be assigned to a combination of three EC positions (EC 5.3.1.9) in glycolysis, in Figure 2, hence, for PGI1 we have \( 2^3 - 1 = 7 \) possible states. Another three genes, HXK1, HXK2 and GLK1, can be assigned to a combination of two positions (EC 2.7.1.1) in glycolysis in Figure 2. Hence, for each we have \( 2^2 - 1 = 3 \) possible states. For the remaining 12 we have only 2 alternatives. Thus, the overall number of combinations is equal to \( 7 \times 3 \times 3 \times 2^{12} = 774,144 \).

The highest acquired fitness in the exhaustive search is 18.5, for just two allocations, while the overall fitness has a mean of 6,774, a minimum of 3.007 and standard deviation of 1.690. The hill climbing search converges to an allocation of fitness equal to 18.5 in just about 200 iterations. This is a huge improvement in efficiency, which would be immense in a larger search setting, e.g. one consisting of more modules or an entire metabolic network. For example 100 expressed genes, even in the simplest case where there are only two alternative allocations for each would require exhaustive search of more than \( 2^{100} \) possible allocations which is not feasible.

6. Conclusions

We have shown that the expression of a multi-membership gene in a module, in a given instance, may not reflect the contribution of the product of that gene to that particular module. For example, up-regulation of such a gene may be due to activation of another module where its protein product participates. Hence, the expression behaviour of multi-membership genes is not a clear indication of the state of expression of gene members of a particular module. The comparatively more frequent expression of such genes than the one of single membership genes and the analysis of the cumulative binomial probabilities discussed here support this hypothesis.

Furthermore, we propose a heuristic search methodology to deal with this issue and correctly identify the state of activation of individual modules. Overall, the method seems capable of successfully differentiating between activation and repression of the glycolysis and gluconeogenesis modules. Importantly, the search setting is such that the algorithm takes into account the competitive nature of the two modules and the topology of the network. That is, the EC positions shared by the modules and those unique to each module, as well as the genes and the positions where they participate are considered and assignment of expressed genes to both modules is disallowed. Rather than simply providing us with a list of expressed genes, the method reveals the precise state of activity in the pathway as a whole, thus providing us with information regarding the direction of the reactions taking place, in this case the degradation or synthesis of glucose.

Here, the analysis is confined to the KEGG glycolysis/gluconeogenesis pathway, consisting of the two competing modules, centred on experiments with sufficient accompanying information, subjected to statistical testing to be confident that the glycolysis/gluconeogenesis pathway is severely affected by the experimental conditions. Naturally, the methodology needs to be extended to the entire metabolic network and it would be interesting to examine the results produced when applied to organisms with more sophisticated biochemistry than Saccharomyces cerevisiae. Organisms higher in the evolutionary chain provide more complicated networks with larger number of genes and pathways, as well as interconnections. The methodology seems an ideal candidate approach, for the identification of the most likely flow taking place in the metabolic network of a cell.

More sophisticated search approaches such as simulated annealing and genetic algorithms are worth exploring in cases of larger search space. This requires detailed preparation of the search setting, identification of the modules constituting each pathway, their interconnectivity and nature. As wet lab experimentation improves pathway knowledge, we may be able to construct this setting in greater detail.

Acknowledgments

This research was partially supported by the PhD studentship scheme of the School of Information Systems, Computing and Mathematics, Brunel University, UK and the EPSRC (grant: EP/H019685/1).
References