Mining pathway signatures from microarray data and relevant biological knowledge

Eleftherios Panteris *, Stephen Swift, Annette Payne, Xiaohui Liu

School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge, Middlesex UB8 3PH, UK

Received 21 September 2006
Available online 22 February 2007

Abstract

High-throughput technologies such as DNA microarray are in the process of revolutionising the way modern biological research is being done. Bioinformatics tools are becoming increasingly important to assist biomedical scientists in their quest in understanding complex biological processes. Gene expression analysis has attracted a large amount of attention over the last few years mostly in the form of algorithms, exploring cluster and regulatory relationships among genes of interest, and programs that try to display the multidimensional microarray data in appropriate formats so that they make biological sense. To reduce the dimensionality of microarray data and make the corresponding analysis more biologically relevant, in this paper we propose a biologically-led approach to biochemical pathway analysis using microarray data and relevant biological knowledge. The method selects a subset of genes for each pathway that describes the behaviour of the pathway at a given experimental condition, and transforms them into pathway signatures. The metabolic pathways of Escherichia coli are used as a case study.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Microarrays; Metabolic pathways; Bioinformatics; Data mining; Pathway analysis; Systems biology

1. Introduction

Microarrays have revolutionised biology as a science and have ignited the bioinformatics revolution. The large amount of data produced by microarray experiments requires constant computational support in order to be of any use. There has been a surge of activity since the early days of bioinformatics to provide tools for the analysis of microarray datasets using clustering techniques and later on to describe the findings by applying modelling techniques [1,2].

Motivated by the need in the bioinformatics area to bridge the gap of communications between biology and computer science [3–5] a new approach to pathway analysis called Signature mining has been proposed. Signature mining provides an analysis framework for microarray data particularly developed for pathway analysis and more specifically for metabolic pathways. A pathway ‘signature’ is a summary measure of the relative expression of a pathway in an experiment. It is created by taking a microarray dataset and then overlaying it with metabolic pathway information. By comparing gene profiles across many different experimental conditions, the most active genes per pathway are found [6]. The combination of these active genes produces a single summary measure for a pathway, which is used to describe the state of that pathway, i.e. its level of activity in any given experimental condition.

Common techniques currently used in the field involve the use of a single set of experiments and a metabolic database, which is superimposed onto the data, graphically portraying the expression per experiment of each gene of the pathway, leaving the researcher to decipher if the pathway is active and to quantify the activity, based on the expressions of a small number of genes per pathway. This ignores the fact that many, sometimes most, genes can be members of several pathways simultaneously. Pathway lists such as KEGG [7] and EcoCyc [8] are verified and composed based

---

* Corresponding author.
E-mail address: eleftherios.panteris@brunel.ac.uk (E. Panteris).

doi:10.1016/j.jbi.2007.01.004
on experimental conclusions from the literature, and clearly demonstrate that most genes are present in more than one pathway, some in as many as seven or more pathways.

The paper begins by describing the current methods used in the area and outlines the reasons why we choose a different approach in Section 2. Section 3 details the proposed algorithm for obtaining pathway signatures. Section 4 provides the experimental results for the pathway signatures on several well known metabolic pathways. Section 5 summarises what has been achieved and suggests future directions for this research.

2. Motivation

Current pathway analysis methods of expression data, which include all the current clustering techniques, require all the genes of a pathway to be taken into account, and may lead to the erroneous conclusion that the activity of a pathway has remained unchanged. For example, if more genes in a pathway are transcriptionally dormant than transcriptionally active, the more numerous dormant ones mask the true picture of a change in the activity of that pathway.

Presently pathway regulation information is not explicit, i.e. it does not specify if pathway A is categorically up regulated in experiment X by Y amount. Several attempts that have been made in the literature include averaging gene expression values of all the genes in the pathway or just overlaying the gene expressions over the pathway outline from pathway databases like KEGG [7] and EcoCyc [8], which is common in most software tools available, such as GenMAPP [9] Pathfinder [10], GScope [11] and Gene-Net [12] to mention a few. These methods are not fully addressing the issue; overlaying microarray data over pathway maps from the KEGG database utilises visualisation techniques to provide structural representation to the data but with no added information on the pathway regulation.

Furthermore, using the average expression of all the genes in a pathway works only on the assumption that the genes are equally responsible for the flow and rate of the pathway something that is not true, especially if we take into account the multiple memberships most genes have, as our method does. Also by using the real or absolute gene expression values to produce an average pathway expression it is assumed that the pathway is the sum of its parts, that is to say the resulting value is of a quantitative not a qualitative description. This is a mistake that has led to wrong conclusions before, as Ma and Zeng showed in their paper [13], and we believe would happen again.

We believe that averaging pathway genes is biologically inaccurate because it implies that the genes contribute equally to the pathway workload and flow, and more importantly, that gene expression is translated directly into a stoichiometric related metric, which is not the case. Gene expression is an indication of behaviour of the cellular machinery, active or dormant, and cannot be used as stoichiometric measurements.

Another method that tries to provide answers in pathways analysis from the clustering point of view by trying to match clusters of genes with functional similarities to pathways is QPACA [14]. They are relying on the observation that only some experiments within a dataset are relevant to a specific pathway and use this experiment to recognize if subsets of genes are part of the same pathway and further predict additional pathway members.

There are several assumptions here that are not in tune with biological reasoning. It might well be that a pathway is more active in a specific experiment, but that does not mean that by choosing only that single experiment we can draw conclusions from the measured gene expressions of the genes as to the pathway they are members of and new potential members to that pathway. It is well documented that microarray experiments often have artefacts and false expression rates; the technology is not advanced and accurate enough to ensure minimum error in both the manufacturing and analysis of microarrays [1,2].

It is statistically reasonable and expected to use replicated experiments to ensure validity of measurement but due to well documented costs and time constrains most datasets do not follow exactly these guidelines. Using a single experiment do not give the flexibility to check the behaviour of the gene across different experimental conditions and the activities measured could be potentially inaccurate. Furthermore, the methods of this kind does not actually provide an answer to the question of pathway behaviour since they are only interested in the structure and discovery of new pathways which is exciting in itself but not practical, considering the fact that the notion of pathways is constantly challenged in the systems biology area [3–5].

3. The methodology: from signature genes to pathway signatures

Biochemical pathways are a rather eclectic collection of genes that are diverse individually but function as unit. The diversity of the genes in most pathways are profound, and genes that code for proteins residing in the membrane of the cell can be in the same pathway with genes that code for transcription factors residing deep into the cell nucleus [15]. Since pathways genes are diverse, they do not necessarily have the same mode of action or response. That is to say, a transmembrane protein will respond differently due to well documented costs and time constrains most datasets do not follow exactly these guidelines. Using a single experiment do not give the flexibility to check the behaviour of the gene across different experimental conditions and the activities measured could be potentially inaccurate. Furthermore, the methods of this kind does not actually provide an answer to the question of pathway behaviour since they are only interested in the structure and discovery of new pathways which is exciting in itself but not practical, considering the fact that the notion of pathways is constantly challenged in the systems biology area [3–5].

Utilising knowledge about biochemical pathways and their components, we tried to produce a practical picture of the behaviour of the metabolism of an organism based on microarray data and pathway data from major databases like KEGG [7]. By collecting numerous experiments from a given organism, *Escherichia coli* in this instance, for distinct environmental conditions and treatments, and then combining it with well-established pathway information
about genes and their biological contribution, we choose a subset of genes from each pathway, the signature genes, which are used to describe the behaviour of that pathway under the given condition.

The signature genes for each pathway are a unique subset from the genes of a pathway that can be monitored in any given microarray experiment to illustrate that pathway’s behaviour. The signature genes are the ‘true’ expression indicators of the pathway. They are the most active genes; hence they are the most sensitive part of the pathway responding to external stimuli. Changes in the environmental conditions affect signature genes in such a way as to alter their expression in the cell more often. We believe that these are the genes affecting the function of the pathway in a given experiment, so by monitoring them, and only them, we can label the pathway as up-regulated, down-regulated or unchanged in the respective experiment. The rest of the genes that constitute the pathway are transcriptionally dormant, meaning, they do not show any fluctuation in expression across experiments, i.e. their transcript levels remaining constant throughout a given condition. By contrast, the expression levels of active genes change often throughout a given condition or between different experimental conditions. We believe that transcriptionally dormant genes are not essential to the regulation of the pathway but rather they provide the infrastructure that is the structural network the pathway relies upon to function.

Monitoring the signature genes of a pathway in all subsequent microarray experimental data would provide an immediate description of the behaviour of the pathway and subsequently of the whole organism in a global pathway/signature network.

Our key emphasis lies on the utilisation of pathway knowledge to group all the scattered genes in a microarray dataset as pathways and monitor the pathway’s behaviour as a whole, rather than genes individually. It is a different concept that aims to help biologists in pathway analysis, by portraying microarray data in a pathway-orientated view, with genes grouped not only by expression similarity but also biologically.

3.1. The signature mining algorithm

In this section we present a framework that utilises microarray data from genes with fluctuating expression levels to describe the state of the biochemical pathway they belong to, at any given experimental condition. Our proposed method extends the work presented at the IDAMAP workshop [6] for which the following algorithm was suggested to find the best selection of genes in each pathway that represent that pathway’s behaviour. This can be very problematic because each gene can be a member of several pathways and we needed to find a way to choose genes that represent each metabolic pathway in a dataset. Essentially we tried to find a way to move genes from one pathway to another based on their similarity of expression for the whole of the 51 experiments used (not just one experiment).

Initially we opted for an algorithm with a hill climbing step [18] described below.

3.1.1. Hill climbing step

Let G be the set of n genes, \( G = \{1, \ldots, n\} \), let \( X \in \mathbb{R}^{n \times T} \) be the n by T gene expression matrix for the n genes where the \( i \)th row of \( X \), \( x_i \), is the gene expression profile for gene \( i \). \( x_{ij} \) is defined as the \( j \)th element of the vector \( x_i \). Let the pathway list \( P \) be a list of m \( >0 \) lists where \( p_i \subseteq G \) is the \( i \)th element of \( P \), where \( |p_i| > 0 \). A signature \( s_i \) of a pathway \( p_i \) is defined as \( s_i \subseteq p_i \), where \( k_i \) is the size of signature \( s_i \). The list of signatures is denoted as \( S \), where \( |S| = m. s_{ij} \) is defined as the \( j \)th element of the list \( s_i \). How close two expression profiles \( a \) and \( b \) are, is defined as follows:

\[
d(a, b) = \sqrt{\sum_{i=1}^{T} (x_{ai} - x_{bi})^2} \tag{1}
\]

\[
D \in \mathbb{R}^{n \times n}, \quad \text{where} \quad D_{ij} = d(i, j) \tag{2}
\]

The \( n \) by \( n \) symmetric matrix \( D \) contains all of the pair-wise similarities between genes. Note that the larger \( d(a, b) \) is, the more dissimilar the genes \( a \) and \( b \) are.

How close together the genes within a signature are is defined as follows:

\[
FS(s_i) = \sum_{a=1}^{k_i} \sum_{b=a+1}^{k_i} d(s_{ia}, s_{ib}) \tag{3}
\]

This is the sum of all pair-wise differences between the elements of a signature.

Eq. (4) represents how well fitted the signatures are, and Eq. (5) represents how many genes have been allocated from each pathway. To ‘mine’ the signatures for each pathway we need to find a set \( S \) where \( F_1 \) is minimised and \( F_2 \) is maximised:

\[
F_1 = \sum_{i=1}^{m} FS(s_i) \tag{4}
\]

\[
F_2 = \sum_{i=1}^{m} k_i \tag{5}
\]

\[
F_3 = \frac{F_1}{F_2} \tag{6}
\]

The algorithm fitness \( F_3 \) is represented in Eq. (6) and needs to be minimised for the optimum solution, i.e. the smallest signature possible with the genes best describing the pathway.

The signature mining algorithm takes as input a Euclidean distance comparison matrix of all the genes from all the pathways, and a pathway list of lists from KEGG [7] of all the pathways and their genes. To mine the appropriate genes for each signature, we decided to randomly remove or replace a gene from a pathway and use a hill climbing technique to evaluate the solution. The evaluation is based on a similarity and a size function, requiring minimisation of their fraction to progress.
3.1.2. Simulated annealing improvement

Since hill climbing is prone to local optima problems, we recently improved the algorithm by transforming the hill climbing step [17] into a simulated annealing one [19] with visible improvements on the speed and performance of the process along with comparisons between the two forms of the algorithm [20].

The simulated annealing step is defined with the equations below. In Eq. (7) the probability that a worst solution is accepted is related to the difference between the solutions $\Delta f$ and the starting temperature $b_0$. The probability of a worse solution being accepted at each iteration reduces as the temperature cools (tends towards zero), as used in [21].

$$p_r = e^{-\Delta f / b_0}$$

$$\Delta f = f(\text{old}) - f(\text{new})$$

$$\varepsilon = b_0 c^{\text{iter}}$$

$$c = \left( \frac{\varepsilon}{b_0} \right)^{\frac{1}{\text{iter}}}$$

Since it is not possible to run the algorithm infinitely we choose the minimum temperature $\varepsilon$ Eq. (8) and in turn this helps calculate the decay constant $c$ Eq. (9) by which the probability of accepting a worse solution is reduced in each iteration. $b_0 = 1000$, and $\varepsilon = 0.01$ were used as starting values. The algorithm has a running time of 14 minutes in MATLAB with $10^5$ iterations.

3.2. The dataset

The data are from the Gene Expression Omnibus (GEO) data repository at NCBI. The E. coli global experiments represent three different experimental conditions in 51 time course experiments in total and contain the majority of the known genes of the organism. The conditions are analysis of changes in gene expression elicited by perturbations of tryptophan metabolism using strains with mutations that affect tryptophan metabolism, analysis of the progression of chromosomal replication forks in synchronized cells, and time course of UV-responsive genes and their role in cellular recovery [16]. The experimental data were normalised to standard deviation of 1 and mean of 0 so that they can be compared together. No further normalisation was necessary since the data were already normalised to log ratios when they were released in GEO [16].

The genes are chosen according to their variability in expression and have to be above a certain global threshold which is empirically defined, as used in microarray analysis [1,17] to be considered as statistically significant. The threshold is empirically selected depending on the dataset used and is considered for each time point independently and the selection process is repeated for every experiment. The E. coli files were taken from the KEGG portal [7]. By combining the two, a list of important genes was assembled and these were used as the input to the algorithm.

As we wanted to get information about metabolic pathways through the genes we focused on metabolic related genes. Using the pathway information available from KEGG, we filtered for genes that were on the metabolic pathways only. This left 2220 genes out of about 4500 genes in total present in the database. From these 2220 genes, we further removed the genes that were under the threshold of 1.4 we used. The 1430 genes left in our data, were the genes that, more than once in the 51 experiments their absolute expression was higher than the threshold. We used these genes to ‘mine’ our signature genes.

For the current statistical analysis the first experimental condition which deals with tryptophan starvation is used. The tryptophan starvation part of the dataset was used for the statistical analysis of the pathway signatures. The number of experiments was 18, from which the first 8 were experiments without tryptophan, with the remaining being experiments in which tryptophan was slowly added to the growth media. The other two conditions were not included in the calculations for the correlation coefficients since correlation between pathways is only biologically relevant only in comparable conditions.

4. Statistical analysis of signature results

4.1. Signature genes

The two versions of the algorithm have been used to mine the signature genes from the phenylalanine, tyrosine and tryptophan biosynthesis pathway, genes with known behaviour from the experimental work of Khodursky et al. [22]. The phenylalanine, tyrosine and tryptophan biosynthesis pathway includes genes from the biosynthesis of these three amino acids. These three processes are grouped together in the KEGG database due to the chemical similarity these amino acids have, thus the pathway contains 26 genes from all three processes.

Both algorithms gave the same genes from the tryptophan pathway as signature genes, B1260 and B1261 [21]. Both were successful in identifying the key genes in the pathway that are able to represent the behaviour shown in Khodursky et al. [22]. The simulated annealing step was, as expected, more accurate at pinpointing the genes with the correct profile by omitting genes with non-relevant profiles than the hill climbing version where the signature genes were included [21].

The signature genes were ‘mined’ from the GEO dataset and applied to the Khodursky et al. [22] dataset. This is an important fact to stress, since that shows that we can find genes that are controlling the expression of the pathway using different datasets and then use only these genes to monitor the experiment at hand. The more extensive the mining dataset the more precise the pinpointing of the signature genes of the pathways will become.

4.2. The pathway signatures

Here we show how the algorithm can be used on a large scale where it is applied to the entirety of the metabolic
pathways of *E. coli*, and the corresponding datasets are described in the next section. The goal is to extract pathway regulation information in a qualitative way. This is done by taking into account the gene expression values of all the *signature* genes we mined for a pathway, and for each experiment/time point we choose the one that is, in absolute terms, the highest value of gene expression amongst them as the representative of the pathway’s behavioural expression.

We believe that a signature gene with the absolute highest value is the bottleneck, the rate controller of the pathway, thus by employing its expression we want to show that the pathway behaves similarly, i.e. it is up or down regulated at this point. The actual expression value is not used as a quantitative measurement of activity because as mentioned before we do not believe it makes biological sense to directly connect gene expression with pathway stoichiometry and rate of production.

Our value shows that the pathway is regulated up or down relatively to the other pathways in the same experiment since the data is already normalised for comparison. Thus this value we use as a summary measure we term *pathway signature* and we use it across the experimental condition to give us the relative behaviour of each pathway for the duration of the experiment.

To validate the *pathway signatures* from our framework, we chose a variety of pathways that are biochemically opposite in metabolism and plotted the *pathway signature* for each experiment in the aforementioned datasets.

The results that follow demonstrate that *pathway signatures* are valid biological representations of the pathway behaviour. We took the biosynthesis and degradation pathways in metabolism of specific molecules and wanted to see if their *pathway signatures* are as distinct and contra-regulated as their pathways clearly are [15].

### 4.2.1. Phenylalanine, tyrosine and tryptophan, metabolic and biosynthetic correlations

Biological knowledge dictates that biosynthetic and catabolic pathways are distinct [15], i.e. they behave in a contra manner to their counter-partners, so when the synthesis of a molecule is up-regulated, and the degradation of the substance is down-regulated.

As with previous work [6,21] the phenylalanine, tyrosine and tryptophan biosynthesis pathway was initially chosen, as defined in the KEGG pathway database, and it is further used to illustrate the biological validity of the method.

Khodursky et al. [22] have done microarray experiments of *E. coli* under tryptophan starvation and by starving the organism in their experiments; they monitored the activation of the aforementioned pathway. Fig. 1 above illustrates the fact by showing how differently the biosynthesis and the metabolism of tryptophan behave as shown by their *pathway signatures*. In this experimental dataset, *E. coli* cells were grown without tryptophan so they have to produce it. As a consequence the metabolism of tryptophan and the other two amino acids that are made by the pathway suppressed while the production is activated. The behaviour shown by the *pathway signatures* is expected and has been well documented notably in Khodursky et al. [22].

To further strengthen the point Table 1 gives the Pearson correlation values of the three *pathway signatures* and Table 2 gives the Kendall’s tau b and Spearman correlations as well. The correlations are all significant at 1% level and are 2-tailed. The Pearson correlations show the linear relationship these pathways have, i.e. the direct influence one has to the other.

We also chose the Spearman’s and Kendall’s tau b correlation coefficients to shown the non-linear relationships between the pathways. As these are metabolic pathways, which are in essence a very large network, many pathways are influenced by proxy from others, and these two correlation coefficients are one way to measure this influence.

These statistical values show that the correlations between the *pathway signatures* of these metabolic processes are in tune with the biological roles of the pathways. The phenylalanine, tyrosine and tryptophan biosynthesis pathway shows a highly linear, negative relationship with all three respective metabolic pathways and with a very high significance. Also bearing in mind that Spearman’s and Kendall’s tau b show non-linear relationships, there is strong non-linear negative correlation between the pathways suggesting that the pathways are in a reciprocal network and not just in a linear path, which describes the actual relationship of these pathways, evident in the KEGG database [7].

This is one example of how the *pathway signature* can be used as a valid, biological representation of a metabolic pathway and how it can closely depict the real world relationship the metabolic pathway has with other metabolic pathways in the same condition.

### 4.2.2. Statistical analysis of some representative *E. coli* metabolic pathway signatures

The same statistical analysis was conducted for all available pathways, 103 in total. Below are some representative metabolic pathways where they were chosen for their importance in metabolism and the fact that they have well documented connections and relationships with other pathways.

The pathways that are almost synonymous to metabolism are of course glycolysis/gluconeogenesis, citrate cycle (TCA), fatty acid biosynthesis and oxidative phosphorylation [15], see Fig. 2. Using their *pathway signatures* for further statistical analysis the correlations show the strength of their relationships both linearly and non-linearly. It is interesting to note how significant the correlations both linear and non-linear are, see Tables 3 and 4. Glycolysis shows a higher and more significant non linear than linear correlation with the TCA cycle because pyruvate will eventually end up in the TCA cycle as a substrate but not directly [15].

We can further see that the *pathway signatures* are indeed a valid and easy pathway representation by showing...
the out of phase co-regulation of three related pathways; the purine and pyrimidine pathways together with the pentose phosphate pathway and glutamate metabolism pathway signatures in Fig. 3. These pathways are the metabolic waste removers of the organism all related to nitrogen metabolism and the removal of urea [15]. Tables 5 and 6 show again their statistical correlations, in all accounts the correlation values are at statistically significant levels. Note especially the non linear correlations that are indeed of very high significance level between the pentose phosphate and the glutamate pathway although their linear relationship is not significant enough.

Glutamate has a high linear relationship with pyrimidine metabolism which in turn has a high linear relationship with the pentose phosphate pathway, which explains the non-linear relationship the correlations are showing is a valid one and also goes some way to explain the similarity of behaviour of the pathways, as seen in Fig. 3.

---

**Table 1**
The Pearson correlation coefficient of phenylalanine, tyrosine and tryptophan metabolic and phenylalanine, tyrosine and tryptophan biosynthesis (PTT) pathway signatures

<table>
<thead>
<tr>
<th></th>
<th>Tyrosine metabolism</th>
<th>Phenylalanine metabolism</th>
<th>Tryptophan metabolism</th>
<th>PTT biosynthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyrosine metabolism</td>
<td>1</td>
<td>.905*</td>
<td>.822*</td>
<td>−.882*</td>
</tr>
<tr>
<td>Phenylalanine metabolism</td>
<td>.905*</td>
<td>1</td>
<td>.919*</td>
<td>−.987*</td>
</tr>
<tr>
<td>Tryptophan metabolism</td>
<td>.822*</td>
<td>.919*</td>
<td>1</td>
<td>−.896*</td>
</tr>
<tr>
<td>PTT biosynthesis</td>
<td>−.882*</td>
<td>−.987*</td>
<td>−.895*</td>
<td>1</td>
</tr>
</tbody>
</table>

* Correlation is significant at the 1% level (2-tailed).

**Table 2**
Nonparametric correlation coefficients of phenylalanine, tyrosine and tryptophan metabolic and phenylalanine, tyrosine and tryptophan (PTT) biosynthesis pathway signatures

<table>
<thead>
<tr>
<th></th>
<th>Tyrosine metabolism</th>
<th>Phenylalanine metabolism</th>
<th>Tryptophan metabolism</th>
<th>PTT biosynthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyrosine metabolism</td>
<td>1</td>
<td>.752*</td>
<td>.516*</td>
<td>−.569*</td>
</tr>
<tr>
<td>Phenylalanine metabolism</td>
<td>.752*</td>
<td>1</td>
<td>.686*</td>
<td>−.582*</td>
</tr>
<tr>
<td>Tryptophan metabolism</td>
<td>.516*</td>
<td>.686*</td>
<td>1</td>
<td>−.608*</td>
</tr>
<tr>
<td>PTT biosynthesis</td>
<td>−.569*</td>
<td>−.582*</td>
<td>−.608*</td>
<td>1</td>
</tr>
</tbody>
</table>

**Kendall’s tau b**

<table>
<thead>
<tr>
<th></th>
<th>Tyrosine metabolism</th>
<th>Phenylalanine metabolism</th>
<th>Tryptophan metabolism</th>
<th>PTT biosynthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyrosine metabolism</td>
<td>1</td>
<td>.903*</td>
<td>.765*</td>
<td>−.798*</td>
</tr>
<tr>
<td>Phenylalanine metabolism</td>
<td>.903*</td>
<td>1</td>
<td>.860*</td>
<td>−.796*</td>
</tr>
<tr>
<td>Tryptophan metabolism</td>
<td>.765*</td>
<td>.860*</td>
<td>1</td>
<td>−.818*</td>
</tr>
<tr>
<td>PTT biosynthesis</td>
<td>−.798*</td>
<td>−.796*</td>
<td>−.818*</td>
<td>1</td>
</tr>
</tbody>
</table>

* Correlation is significant at the 1% level (2-tailed).
This is a common behavioural motif between biosynthesis and metabolic pathways as we know [15] and there are numerous examples to further illustrate this.

Table 3
The Pearson correlation coefficients of the glycolysis/gluconeogenesis, TCA, fatty acid biosynthesis and oxidative phosphorylation pathway signatures

<table>
<thead>
<tr>
<th>Pathway Signatures</th>
<th>Glycolysis/gluconeogenesis</th>
<th>Citrate cycle (TCA cycle)</th>
<th>Fatty acid biosynthesis (path 1)</th>
<th>Oxidative phosphorylation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycolysis/gluconeogenesis</td>
<td>1</td>
<td>.547&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.633&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.612&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Citrate cycle (TCA cycle)</td>
<td>.547&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1</td>
<td>.942&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.897&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Fatty acid biosynthesis (path 1)</td>
<td>.633&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.942&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
<td>.931&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Oxidative phosphorylation</td>
<td>.612&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.897&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.931&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
</tr>
</tbody>
</table>

<sup>a</sup> Correlation is significant at the 1% level (2-tailed).

<sup>b</sup> Correlation is significant at the 5% level (2-tailed).

Table 4
The non-parametric correlation coefficients of the glycolysis/gluconeogenesis, TCA, fatty acid biosynthesis and oxidative phosphorylation pathway signatures

<table>
<thead>
<tr>
<th>Pathway Signatures</th>
<th>Glycolysis/gluconeogenesis</th>
<th>Citrate cycle (TCA cycle)</th>
<th>Fatty acid biosynthesis (path 1)</th>
<th>Oxidative phosphorylation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kendall’s tau b</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycolysis/gluconeogenesis</td>
<td>1</td>
<td>.490&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.556&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.490&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Citrate cycle (TCA cycle)</td>
<td>.490&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
<td>.882&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.817&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Fatty acid biosynthesis (path 1)</td>
<td>.556&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.882&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
<td>.725&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Oxidative phosphorylation</td>
<td>.490&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.817&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.725&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
</tr>
<tr>
<td>Spearman’s rho</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycolysis/gluconeogenesis</td>
<td>1</td>
<td>.662&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.744&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.703&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Citrate cycle (TCA cycle)</td>
<td>.662&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
<td>.969&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.915&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Fatty acid biosynthesis (path 1)</td>
<td>.744&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.969&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
<td>.876&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Oxidative phosphorylation</td>
<td>.703&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.915&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.876&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1</td>
</tr>
</tbody>
</table>

<sup>a</sup> Correlation is significant at the 1% level (2-tailed).

Fig. 2. The pathway signatures of glycolysis, TCA cycle, fatty acid biosynthesis and oxidative phosphorylation.

This is a common behavioural motif between biosynthesis and metabolic pathways as we know [15] and there are numerous examples to further illustrate this.

The statistical analysis of the whole dataset has shown a large number of correlations between the metabolic pathways that for the most pathways are readily verifiable in the KEGGS database. Further results were not shown due to space limitations; however they all show that the pathway signatures are indeed a valid way of portraying metabolic pathways.
5. Conclusions and future work

5.1. Pathway signatures and systems biology

We have shown that the signature mining framework is a valuable tool for pathway analysis in biomedical research due to its ability to portray all the metabolic pathways of an organism in a single experiment providing at the same time with definitive behavioural information. This is done not only by finding the relevant genes in each pathway but also by providing a single relative value for each pathway that shows if the pathway is active, as well as by how much compared to the others and whether it is up or down regulation at the same time.

The researcher is able to see which pathways are affected in each experiment and how, following a specific treatment, being a drug or a substance, the organism responds. This we believe allows the scientist to make
valuable hypotheses on the way the organism reacts to a specific course of treatment and show potential side effects by highlighting the pathways responsive to the treatment for further tests.

5.2. Pathway signatures

Our pathway signature has several advantages over other methods, the most prominent being the definite answer to the question “Is a pathway active and if yes is it up or down regulated in a given experiment?” This is very important for the scientist because it can provide a definite answer for all the pathways in a set of experiments.

Another advantage, is that since we extracted a singular value that describes the behaviour (we can term it relative expression) of a pathway, we can apply common gene expression techniques to further aid pathway analysis, i.e. pathway signatures can be compared to each other to see potential pathway coordinated behaviour that can mean co-regulation or inverse relationships in pathways. In essence our framework allows us to analyse microarray data in modular ways and explore inter-pathway relationships in a systems biology mindset.

5.3. Future work

Future work includes the exploration of clustering techniques in order to group the pathways in meaningfully functional clusters to allow for systemic exploration of metabolism, and identification of new pathway members from the large number of genes with unknown functions. Also, a more formal analysis of the algorithm is planned to establish areas that need optimisation. Furthermore, applications of the algorithm will not be restricted to E. coli but to other organisms with specific pharmaceutical concerns and ultimately to human data, with a continuation of the framework steps to include gene networks and interactions with protein–protein networks, in order to offer a valid solution in that area of systems biology. The algorithm m-files and links to the datasets used can be found at http://people.brunel.ac.uk/~csstxhl/IDA/SignaturesJBI.

Acknowledgments

The authors would like to take this opportunity to thank the reviewers for the critical evaluation of the manuscript and the very helpful comments and questions that allowed us to improve it, as well as convey our work and ideas in a refined way.

References