Signature Mining: a Heuristic Approach to Biochemical Pathway Analysis

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

School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge, Middlesex UB 8 3PH, UK
*Contact e-mail Eleftherios.Panteris@brunel.ac.uk
http://www.ida-research.net

Abstract

Microarrays have revolutionised biology, and bioinformatics is now a powerful tool in the hands of biologists. Gene expression analysis has attracted a large amount of attention over the last few years mostly in the form of algorithmic explorations of cluster relationships, and software that try to display the multidimensionality of microarray data in biological relevant formats. In this paper we propose a simple yet effective approach to biochemical pathway analysis based on biological knowledge, to select a subset of genes for each pathway that fully describes the behaviour of the pathway at a given experimental condition in a bid to reduce the dimensionality of microarray data and make the analysis more biologically relevant.

1 Introduction

A new field attempting to describe biology called systems biology is currently emerging trying to depict biology at an organisation level by multidisciplinary research [Aggawal *et al.,* 2003]. Microarrays are an essential member of this multidisciplinary approach and a lot of interest has focused on gene expression analysis. Informatics and computer science are important members of this field with a heavy interest in microarray data analysis and data storage, as well as in distribution and display of data in terms of clustering programs and large databases. Network modelling is also very active trying to describe biochemical pathways and biological processes in general [Huang, 2004].

These multidisciplinary approaches aspire to combine and produce practical descriptive models of biological systems that can be used among others to predict drug response and aid in cancer prevention and treatment.

Analysis of microarray gene expression [Eisen *et al.,* 1998] has attracted a lot of attention over the years mostly the form of algorithmic explorations of cluster relationships, and software that try to display the multidimensionality of microarray data in biological relevant formats [Slonim, 2002]. The multi-dimensionality of the microarray experimental data has made this into a daunting task and there still a lot to be desired from the current work [Claverie, 1999]. Meanwhile, the modelling community has a growing interest in the complexity of biochemical pathways and various modelling methods exist that try to predict how such pathways behave [Papin *et al.,* 2003].

We have taken a simple yet effective approach to pathway analysis using the idea of *signatures* for each pathway. An algorithm based on hill climbing was used to mine for the *signatures* in all the 108 pathways from *E.coli*. The algorithm is effective in finding biologically relevant *signatures* and the results are promising that this is a valid way forward in the field.

The background behind this study is in Section 2, giving the reasons why we used a novel interpretation of biochemical pathways for our problem. Section 3 gives information about the data and their sources, and section 4 describes the *signature* mining process and its algorithm. Section 5 deals with the biological verification of the results. Section 6 summarises the findings and proposes future directions of work.

2 Background

The identification and validation of drug targets depends critically on knowledge of the biochemical pathways in which potential target molecules operate within cells. For this reason, the study of biochemical pathways is the focus of numerous drug discovery researchers and is central to the strategy of many biopharmaceutical and genomic companies.

There is intense research going on in systems biology, with scientists using different methods to solve similar problems [Aggawal *et al.,* 2003]. From the biological point of view, most scientists use methods that offer some but not all of the functionality a biologist would like to have, often with rather complex and time consuming implementations. If pathway analysis and visualisation is going to be performed by biologists alone, it should be done in a straightforward and with few intermediate steps way so scientists can focus on the biological significance of the findings and not the programming implementation of the methods. So far this is not currently available in the research community, and the software available, both public and commercial, do not provide all the functionality they should. [Goesmann *et al.,* 2002; Toyoda *et al.,* 2003; Dahlquist *et al.,* 2002; Kolpakov *et al.,* 1998]

These constraints infiltrate the relationship biologists have with other sciences and computer science in particular. Ma and Zeng have shown in their paper [2003] that modelling biochemical pathways is not straightforward, and mistakes can be made if all parameters are not taken into account.

Utilising biological knowledge about biochemical pathways and their components, this study produces a practical picture of the behaviour of the whole genome of an organism based on microarray data and pathway data from major databases like KEGG [Kanenisha and Goto, 2000]. By collecting numerous experiments from a given organism, *E.coli* was used, for distinct environmental conditions and treatments and then combining it with well-established pathway information about genes and their biological contribution, we choose a sub set of genes from each pathway, a '*signature*', which is used to describe the behaviour of that pathway under the given condition.

A pathway's *signature* is a unique set of genes that can be monitored in any given microarray experiment to illustrate that pathway's behaviour. The *signature* is the collection of the 'true' *expression indicators* from the pathway. They are the most 'expressively active' genes, in the sense that they are the more sensitive part of the pathway, the ones most responsive to external stimuli, i.e., the change in the environmental conditions affects them in such a way as to alter their expression in the cell. The rest of the genes in that pathway are transcriptionally dormant in the sense that they do not respond readily to change, since they form the infrastructure of the pathway in the cell, and as building block they are not sensitive to external stimuli as much.

Pathway analysis methods of expression data currently in use, which include all the recent 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 remain 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.

By monitoring the *signature* of a pathway in all subsequent microarray experimental data we would have an immediate description of the behaviour of the pathway and subsequently of the whole organism in a *global pathway /signature network*. In essence, we aim to reduce the dimensionality of cDNA microarray data to provide a biologically relevant picture of the whole organism immediately, before resorting to clustering methods.

Our emphasis lies on using pathway knowledge to group all the scattered genes in a microarray dataset as pathways and observe 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 representing microarray data in a pathway-orientated view, with genes grouped not only by expression similarity but also biologically.

Furthermore, it offers a simplified view of these pathways by using a specific subset of genes to depict the behaviour in each experiment. This offers new options to biologists who could group or 'cluster' the pathways according to behaviour in an experiment thus, finding interesting connections, not easily observed in gene clustering techniques and visualisations.

3 Data

Gene **E**xpression **O**mnibus data repository at NCBI was the source of the microarray datasets. They come from *E.coli* and represent three different experimental conditions in 51 experiments in total. We exploited the variety of conditions to find the most sensitive genes under these conditions, since the larger the number of experimental conditions and number of experiments, the more fine-tuned the dataset is. There are global cDNA microarray experiments containing the majority of the *E.coli* genes. The experimental data, representing 51 microarray experiments, 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**.

The genes are chosen according to their variability in expression and have to be above a certain empirically defined global threshold, as used in microarray analysis 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 threshold is the statistically significant fold difference between the two copies of the gene in the control and test conditions. Its range is usually between 1 and 2 fold and the researcher chooses an appropriate value depending on the general levels of expression of all the genes in the experiment. [Schena *et al.,* 1996; Dunggan *et al.,* 1999]
The KEGG *E.coli* files were taken from the KEGG portal [Kanenisha and Goto, 2000]. By combing the two, a list of important genes was assembled and these were used as the base of the algorithm.

4 Algorithm

Choosing the best selection of genes in each pathway that represent that pathway's behaviour is challenging because each gene can be a member of several pathways and we needed to find a way to choose genes that represent each pathway out of the 108 of *E.coli*. Basically 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 not just one experiment. We opted for an algorithm with a hill climbing [Michalewicz *et al.,* 1998] step described below.

Let G be the set of n genes, $G = \{1,\ldots,n\}$, let $X \in \Re^{n \times T}$ be the n by T gene expression matrix for the n genes where the ith 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 $|s_i| > 0$. The list of *signature*s is denoted as *S*, where $|S| = m$. s_{ij} is defined as the *j*th element of the list s_i, such that s_i is a subset of the corresponding p_i. How close two expression profiles *a* and *b* are, is given by the Euclidean Distance formally defined in formula 1.

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

$$D \in \Re^{n \times n}, \text{ where } D_{ij} = d(i,j) \quad (2)$$

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:

$$FS(s_i) = \sum_{a=1}^{|s_i|-1} \sum_{b=a+1}^{|s_i|} d(s_{ia}, s_{ib}) \quad (3)$$

This is the sum of all pair-wise differences between the elements of a *signature*. Equation 4 represents how well fitted the *signatures* are, and equation 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:

$$F_1 = \sum_{i=1}^{m} FS(s_i) \quad (4)$$

$$F_2 = \sum_{i=1}^{m} |s_i| \quad (5)$$

$$F_3 = \frac{F_1}{F_2} \quad (6)$$

The algorithm fitness F3 is represented in equation 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 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. The algorithm is described below (algorithm1).

Algorithm 1 *Signature* Mining:

```
(1) INPUT:
    Euclidean Distance Matrix (D)(eq.
    2),Filtered Pathway list of lists
    (PA)(108 pathways long).
(2) ITERATIONS
    For w=1:ITER do
(3) RANDOM SELECTION with REMOVAL or
    REPLACEMENT
    Randomly chose a pathway and randomly
    chose a gene position from the
    pathway.
    If gene is present:
       REMOVE
    Else
       REPLACE gene back to position.
    End
(4)     GET EUCLIDIAN DISTANCES FROM ALL
        PATHWAYS
        For i=1:length of PA do
            OBTAIN all unique distances
            between the pathway genes
            from D for comparison.
        End
(5) GET F1 (SIMILARITY FUNCTION)(eq. 4)
    For i=1:length of PA do
        F1(i)= SUM(distances of all
        genes from P(i))
    End
       Store F1(i)
(6)     GET F2 (SIZE FUNCTION)(eq. 5)
        For i=1:length of PA do
            F2(i)= length of P(i)
        End
            Store F2(i)
(7) GET F3(EVALUATION FUNCTION)(eq. 6)
    For i=1:length of PA do
        F3(i)=SUM(F1(i)/F2(i))
        F3new(w)=F3new(w)+F3(i)
    End
(8)      EVALUATION
         If F3new(w) < F3old
            SET as F3old
         Elseif F3new(w) > F3old
               RESET to previous value
         End
    End
(9) OUTPUT:Signature list for all (108)
    pathways
```

The convergence of the algorithm is shown below (Fig.1), the number of iterations being 20000. The convergence graph shows that the algorithm performs well, by sharply dropping for the first 4000 iterations and then slowly stabilising to the minimum evaluation value possi-

ble, from the 6000 iteration onwards the slope levels up to almost a straight line.

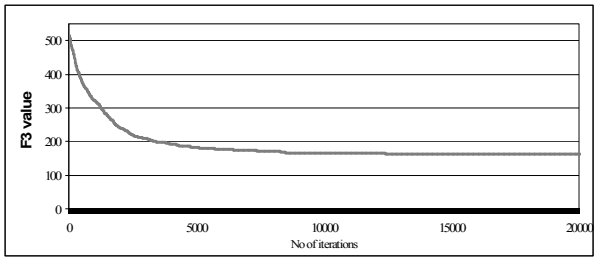

Fig. 1. Convergence plot of the Algorithm. The y axis F3 value refers to the evaluation function F3 (equation 6).

The performance of the algorithm accordingly can be seen from the histograms that represent the distribution of genes per pathway (Fig.2), both before (a) and after (b) the application of the algorithm. By keeping only the *signature* specific genes after using the algorithm, each pathway is reduced to at least 1/3 of its previous size, making it more specific, allowing for easier biological interpretation of the pathway behaviour.

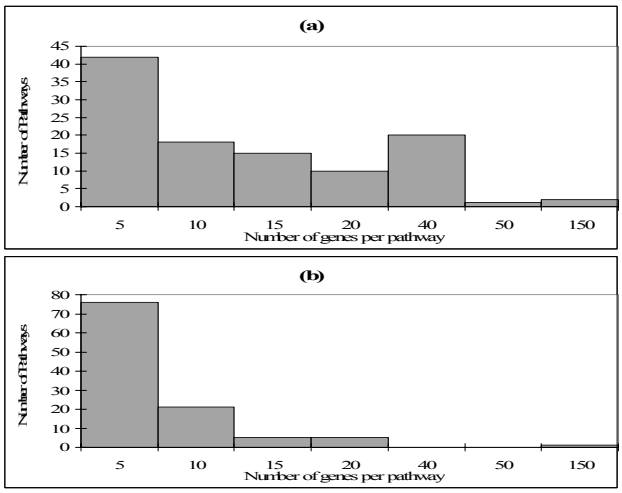

Fig. 2. The distribution of genes per pathway before (a) and after (b) the application of the signature algorithm, with number of pathways and number of genes per pathway.

5 Application

In order to biologically validate our method we chose a pathway from *E.coli*. The Phenylalanine, Tyrosine and Tryptophan biosynthesis pathway, as defined in the KEGG pathway database, was chosen with focus on Tryptophan. The Tryptophan production is regulated from a specific operon that contains five genes, B1260, B1261, B1262, B1263, and B1264.

A microarray experiment of *E.coli* under Tryptophan starvation from Khodursky et al [2000] was used. They observed a very specific response from the Tryptophan operon genes. These genes are activated in the absence of Tryptophan and induce its production. So by starving the organism in their experiments, they monitored the activation of the pathway.

Using the signature mining algorithm we 'mined' a signature for the specific pathway that portrays the behaviour of the pathway according to Khodursky et al [2000]. The importance of the signature lays in the fact that we used the GEO dataset, to find the signature that describes the pathway in the Khodursky et al [2000] dataset.

The Phenylalanine, Tyrosine and Tryptophan biosynthesis pathway includes genes from the biosynthesis of these three amino acids. They are grouped together in the KEGG database due to the chemical similarity these amino acids have.

The pathway contains 26 genes and the signature mining algorithm produced a pathway signature of 6 genes. The 6 genes are from all the three branches of the specific pathway as shown in table 1. Genes B0928 and B2021 are present in both the phenylalanine and tyrosine processes.

Table 1. Distribution of signature genes per branch of the Phenylalanine, tyrosine and tryptophan pathway.

PHENYLALANINE	TYROSINE	TRYPTOPHAN
B1713	B4054	B1260
B0928	*B0928*	B1262
B2021	*B2021*	

Khodursky et al [2000] are interested only in Tryptophan starvation so their dataset contains only the genes B1260 and B1262. The starvation response of *E.coli* is to activate the genes that produce Tryptophan [Khodursky et al., 2000]. The response can be observed in Fig. 3 where the gene expression of the genes that constitute the tryptophan operon is plotted in a six part starvation time course. The organism is placed in an environment without tryptophan at the start of the experiments (see Fig.3) and gene expression measurements are taken at 20 minutes intervals. It is obvious from the graph (Fig 3) that the genes are highly up-regulated moments after the starvation initiation.

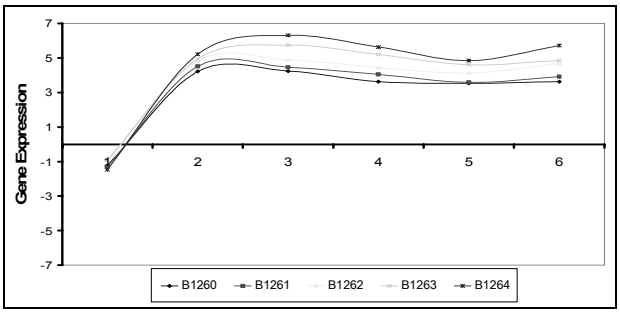

Fig. 3. The Tryptophan operon activation from the Khodursky et al [2000] dataset. See text for further details.

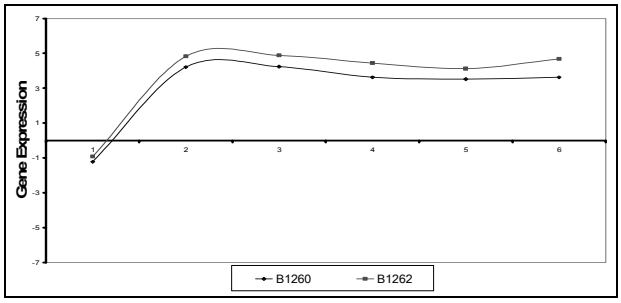

Fig. 4. The Signature genes present in the dataset. The genes describe the activation of the operon during Tryptophan starvation.

Our signature has two out of five genes from the tryptophan operon. As it can be seen from Fig. 4 they are sufficient to portray the behaviour of the pathway during the experiments. As mentioned above briefly, the signature was 'mined' from the **GEO** dataset and applied to the Khodursky *et al* dataset [2000].

This has provided early evidence that signature mining can be an effective way of analysing biochemical pathways because once they are chosen they can be applied across experimental conditions and datasets with ease.

The biological relevance of the *signature* mining algorithm is extensive, especially in the biochemical and pharmaceutical community, since it allows the biologist to observe the behaviour of a specific pathway in a clear and definitive way that does not involve genes that do not affect the pathway's regulation. Its relevance is evident in drug related research. *Signature* mining could help by monitoring the effect of the drug on the whole of the organism by reducing the complexity of the pathways and showing a holistic view of the organism. An up or down regulated behaviour could be easily identified in the signature context and that pathway chosen for further investigation to find the specific genes affected. In essence, it allows the researcher to have an overview of the entire organism processes in an simple and obvious way, easy to understand and use.

6 Conclusions

To conclude, we have shown that a specially selected sub group of genes from a biochemical pathway, a *signature*, is able to depict its behaviour under a given experimental condition. A simple yet effective algorithm based on hill climbing was created to mine for the appropriate genes for each pathway based on the pathway's behaviour across a large set of experiments of varying conditions. The algorithm was able to select *signatures* for all 108 pathways, of which one was used as an example here.

Using biological knowledge in the design of the algorithm was very important as it was the biological verification of the results. The preliminary results have clearly shown this interpretation of biochemical pathways to be an interesting way of describing microarray data used for pathway analysis.

Future work will improve the algorithm run time and functionality. The algorithm needs to be more selective and not prone to be trapped in local optima. Ideally the algorithm should have picked all five genes that form the operon, this maybe due to the common problem of hill climbing algorithms of being easily trapped in local optima and it is currently being addressed using a global search method such as simulated annealing [Kirkpatrick *et al.,* 1983].

Additionally, the algorithm will be a part of a framework for microarray datasets for full exploitation of microarray data in relation to pathway analysis and pharmaceutical research. Application of the algorithm will not be restricted only to *E.coli* but to other organisms with specific pharmaceutical concerns and in due course to human data, with a continuation of the framework steps to include gene networks and interactions with protein-protein networks, offering a solid solution in that area of systems biology.

Acknowledgements

The authors would like to thank the reviewers for their useful insights and comments about the manuscript.

References

[Aggawal *et al.,* 2003] Aggawal, K., Lee, K.H.,; Functional genomics and proteomics as a foundation for systems biology.' Briefings in functional genomics and proteomics Vol 2, No 3, 175-184 2003

[Claverie, 1999] Claverie, J.; Computational methods for the identification of differential and coordinated gene expression. Human Molecular Genetics. Vol 8, No 10 1821-1832 1999

[Dahlquist *et al.,* 2002] Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C., Conklin, B.R.,; GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways, Nature Genetics 31(1):19-20 2002

[Duggan *et al.,* 1999] Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. & Trent, J.; Expression profiling using cDNA microarrays. Nature Genetics. 21, 10-14 1999

[Eisen *et al.,* 1998] Eisen, M., Spellman, P.T., Botstein, D., Brown, P.O,; Clustering Analysis and display of genome-wide expression patterns. PNAS 95. 14863-14868 1998

[Goesmann *et al.,* 2002] Goesmann, A., Haubrock, M., Meyer, F., Kalinowski, J., Giegerich, R.; PathFinder: reconstruction and dynamic visualization of metabolic pathways. Bioinformatics. 18: 124-129 2002

[Huang, 2004] Huang, S.; Back to the biology in systems biology: What can we learn from biomolecular net-

works? Briefings in functional genomics and proteomics. Vol 2, No 4, 279-297 2004

[Kanehisa and Goto, 2000] Kanehisa, M., Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 28, 27-30 2000

[Kirkpatrick *et al.,* 1983] Kirkpatrick, S., Gelatt, Jr., C.D., Vecchi, M.P.; Optimization by Simulated Annealing, Science. 220, No. 4598, 671-680 1983

[Khodursky A., *et al* 2000] Khodursky AB, Peter BJ, Cozzarelli NR, Botstein D, Brown PO, Yanofsky C. DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *E.coli***.** PNAS. USA. Oct 24;97(22):12170-5 2000

[Kolpakov *et al.,* 1998] Kolpakov, F.A., Ananko, E.A., Kolesov, G.B., Kolchanov, N.A.; GeneNet: a gene network database and its automated visualization. Bioinformatics 14: 529-537 1998

[Ma and Zeng, 2003] Ma, H., Zeng, A.,; Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics 19: 270-277 2003

[Michalewicz *et al.,* 1998] Michalewicz, Z., Fogel, D. B.; How To Solve It: Modern Heuristics, Springer, 1998.

[Papin *et al.,* 2003] Papin, J., Price, N., Wiback, S., Fell, D., Palsson, B.; "Metabolic pathways in the post-genome era," Trends in Biochemical Sciences., Vol. 28, 250-258 2003

[Schena *et al.,* 1996] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., Davies, R.W.; Parallel Human Genome Analysis: Microarray-Based Expression Monitoring of 1,000 Genes. PNAS 93. 10614-10619 1996

[Slonim, 2002] Slonim, D. K.; From patterns to pathways: gene expression data analysis comes of age. Nature Genetics. 32, 502 - 508 2002

[Toyoda *et al.,* 2003] Toyoda, T., Mochizuki, Y., Konagaya, A.,; "GSCope: a clipped fisheye viewer effective for highly complicated biomolecular network graphs," Bioinformatics, Vol. 19, 437-438 2003