

MicroPath-A pathway-based pipeline for the comparison of multiple gene expression profiles to identify common biological signatures

Mohsin Khan¹, Chandrasekhar Babu Gorle¹, Ping Wang³, Xiao-Hui Liu², Su-Ling Li¹

¹Molecular Immunology & bioinformatics Group, Microarray Facility, Division of Bio-Sciences, Brunel University, Uxbridge, UB8 3PH, UK; ²Intelligent Data Analysis Group, Department of Information Systems and Computing, Brunel University, Uxbridge, UB8 3PH, UK; ³Immunology Group, Institute of Cell and Molecular Sciences, Barts and London School of Medicine, London, UK. Correspondence should be addressed to Su-Ling Li (su-ling.li@brunel.ac.uk)

Received Jan. 2nd, 2009; revised Feb. 15th, 2009; accepted Mar. 4th, 2009.

ABSTRACT

High throughput gene expression analysis is swiftly becoming the focal point for deciphering molecular mechanisms underlying various different biological questions. Testament to this is the fact that vast volumes of expression profiles are being generated rapidly by scientists worldwide and subsequently stored in publicly available data repositories such as ArrayExpress and the Gene Expression Omnibus (GEO). Such wealth of biological data has motivated biologists to compare expression profiles generated from biologically-related microarray experiments in order to unravel biological mechanisms underlying various states of diseases. However, without the availability of appropriate software and tools, they are compelled to use manual or labour-intensive methods of comparisons. A scrutiny of current literature makes it apparent that there is a soaring need for such bioinformatics tools that cater for the multiple analyses of expression profiles.

In order to contribute towards this need, we have developed an efficient software pipeline for the analysis of multiple gene expression datasets, called MicroPath, which implements three principal functions; 1) it searches for common genes amongst n number of datasets using a number crunching method of comparison as well as applying the principle of permutations and combinations in the form of a search strategy, 2) it extracts gene expression patterns both graphically and statistically, and 3) it streams co-expressed genes to all molecular pathways belonging to KEGG in a live fashion. We subjected MicroPath to several expression datasets generated from our tolerance-related in-house microarray experiments as well as published data and identified a set of 31 candidate genes that were found to be co-expressed across all interesting datasets. Pathway analysis revealed

their putative roles in regulating immune tolerance. MicroPath is freely available to download from: www.1066technologies.co.uk/micropath.

Keywords: Co-Expression Analysis, Microarray, Permutations and Combinations, Multiple Gene Expression Analysis

1. INTRODUCTION

There is a general consensus amongst scientists and researchers that the fundamental asset of microarray technology lies in its inherent ability to produce a global snapshot of the cellular state in the milieu of any given biological question. It is therefore not surprising that microarrays have revolutionised the field of molecular biology by offering an efficient and cost effective medium for biologists to quantify mRNA transcript levels of several thousands of genes concurrently in order to observe specific states of the transcriptome (in response to a particular treatment or specific time point). Owing to this innate faculty to decipher the transcriptome, gene expression profiles pertaining to a wide variety of biological questions are being rapidly generated by scientists worldwide and are deposited and subsequently made accessible through public repositories such as ArrayExpress [1] and the Gene Expression Omnibus [2]. With so much wealth of high throughput biological data made available, biologists have become motivated to utilise these sets of data in an attempt to investigate common regulatory signatures, which may be implicating the transcriptome state across multiple gene expression profiles sharing a similar biological theme. One of the most widely accepted methodologies of comparing expression profiles is based on the assumption that genes across different biological conditions sharing similar expression patterns are likely to be involved in the same biological processes [2], and therefore, may share common regulatory signatures. By using this method of comparison, which is one of the most successful methods to date, coupled with the availability of publicly available data

repositories offering gene expression profiles, biologists have been granted the opportunity to answer complex biological questions pertinent to biological phenomena underlying various different disease states.

To this end, we have developed a novel bioinformatics software pipeline called MicroPath, which specialises in the cross comparison of multiple gene expression datasets and attempts to identify common regulatory signatures from the standpoint of molecular pathway analysis. When one scrutinises current literature relevant to automated solutions of gene expression analysis, it becomes apparent that there is an increasing demand for software applications that offer an efficient pipeline to the analysis of multiple gene expression profiles. Although current meta-analyses studies have been conducted with the purpose of employing statistical techniques to compare cDNA and affymetrix gene expression profiles [3,4,5,6], it cannot be denied that there is a mounting need for this process to be automated. Nevertheless, various approaches/algorithms of statistical nature have already been implemented with the purpose of identifying the most relevant pathways in a given experiment [7,8,9] together with methods such as Gene Set Enrichment Analysis (GSEA), which ranks genes based on the correlations between their expressions and observed phenotypes in the context of biological pathway discoveries [10]. There are also tools available that functionally annotate gene expression data [11,12]. Albeit, it remains infeasible for biologists to cross compare several expression profiles without an automated solution, and hence, they are faced with the labour-intensive task of employing manual methods to carry out their comparisons. MicroPath uses the meta-analytic standard and has been specifically developed to: compare several significantly expressed sets of genes in order to find the intersection of common genes using both number crunching methods as well as the classical permutation and combination principle, extract putative regulatory signatures using both statistical and graph-based approaches and finally, mapping these sub-sets of co-expressed genes to molecular pathways all in the form of a high throughput pipeline.

2. IMPLEMENTATION

The front-end of MicroPath was developed in Visual Basic.Net and Perl, and the database back-end was developed in MySQL. Upon analysing the users input files (gene expression profiles), processed data is displayed intuitively on the graphical user interface, which is equipped with various interactive objects such as charting facilities, buttons, drop-down menus and user input/output dialogues. The interface is also equipped with a function to export processed data into Microsoft excel for further scrutiny and use.

2.1. System Architecture

MicroPath carries out meta-profiling of multiple gene expression datasets using two different approaches.

Firstly, the intersection of common genes is identified across n number of expression profiles, which is then plotted graphically using a simple number crunching exercise. The second approach applies to a situation where an attempt to identify common genes across n number of expression profiles using the aforementioned approach fails due to the absence of common genes across all datasets (this situation is especially common when a large number of expression profiles are compared, which reduces the probability of finding a common gene amongst them). Consequently, MicroPath applies the permutations and combinations mathematical principle to solve this problem (refer to *implementation of meta-analysis strategy* below for details). Once the intersection of a set of common genes has been identified and subsequently displayed on the interface (using either of the above methods), the next stage in the analysis is to extract patterns from the intersection in order to identify common genes that are being expressed in accordance with the biological question. MicroPath offers a semi-automated graph-based approach to achieve this as well as classical statistics to identify the overall correlation of gene expression. Finally, co-expressed genes (common genes that are expressed in accordance to the relevant biological question) are mapped to all molecular pathways known to date in order to reveal their molecular dependencies (refer to **Figure 1** for the complete system architecture).

2.2. Implementation of Meta-analysis Strategy

In theory, an intersection of a sub-set of common genes across multiple gene expression profiles should be easily attainable using simple number crunching methods of comparison. In practice, this is not always the case since the likelihood of identifying genes sharing common accession identifiers decreases as the number of profiles to compare increases. This inverse relationship makes sense both mathematically and biologically. From a biological perspective, regulatory signatures tend to be diluted over entire datasets and as a result, only a proportion of the total number of profiles to compare may actually share common genes. In such a scenario, using a simple method of comparison would break down at some point and no common genes would be reported to the user, although common genes may be present within $n-1$ expression profiles. To prevent potentially interesting biological findings to be hampered at this point in the analysis, we have applied the principle of mathematical combinations to the comparison of multiple gene expression profiles. All possible combinations of comparing n number of datasets with each other are firstly computed using the combination equation:

$$\pi_C = \frac{n!}{R!(n-r)!}$$

Where $n!$ is the factorial of the total number of datasets n , and $r!$ is the factorial of the selected number of datasets to compare when comparing n datasets results in zero common genes, r .

This generates the total number of permutations of comparing datasets (C_r) for given values of n (total number of datasets imported by user) and r (number of

intended datasets used to search for common genes when zero common genes are reported across n datasets) (**Table 1**).

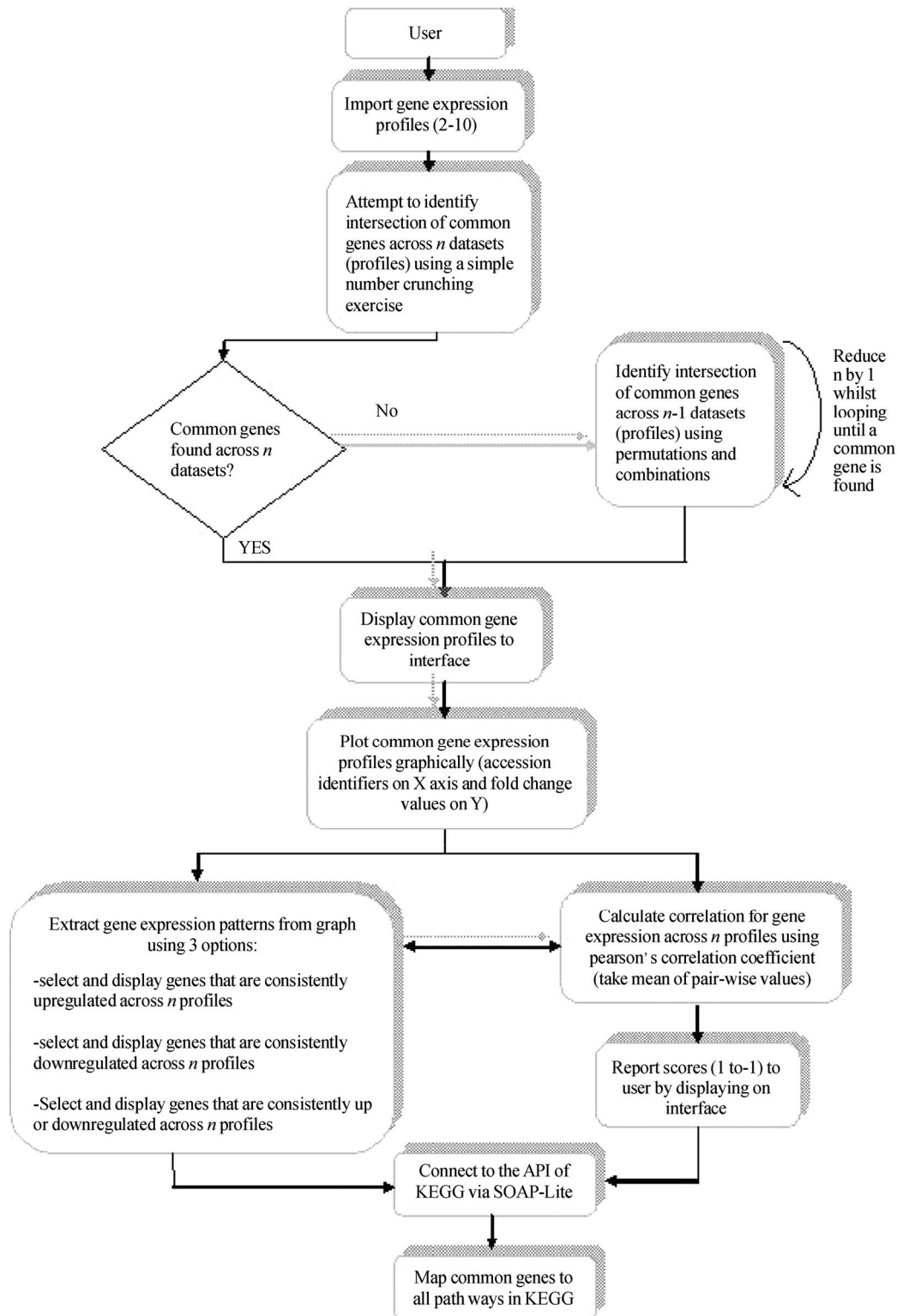


Figure 1. Functions of MicroPath. Users are prompted to import up to 10 gene expression profiles, which are then compared using a direct comparison method. If this method yields zero common genes, MicroPath automatically attempts to identify an intersection of common genes by reducing the search space to $n-1$ datasets using permutations and combinations. This process is continued until at least 1 common gene is reported. Following this, users are provided with a function to search for expression patterns graphically and gene expression correlations are calculated statistically using the pearson's correlation coefficient algorithm. Finally, co-expressed genes are mapped to all molecular pathways of KEGG in a high throughput fashion by automatically accessing its API via SOAP-Lite.

Table 1. Multiple gene expression profile search strategy generated from applying the principle of permutations and combinations. The first column represents the total number of expression datasets, n , that users may import (this is the search space). The second column represents, r , the number of expression datasets to compare if zero common genes are reported to be matched across n datasets. The final column represents the total number of mathematical combinations possible for each given value of n and r .

Total number of expression datasets (n)	Number of intended expression datasets to compare when comparing n datasets yields no results (r)	$n - r$	Total number of combinations of r (Cr)
10	9	1	10
10	8	2	45
10	7	3	120
10	6	4	210
10	5	5	252
10	4	6	210
10	3	7	120
10	2	8	45
9	8	1	9
9	7	2	36
9	6	3	84
9	5	4	126
9	4	5	126
9	3	6	84
9	2	7	36
8	7	1	8
8	6	2	28
8	5	3	56
8	4	4	70
8	3	5	56
8	2	6	28
7	6	1	7
7	5	2	21
7	4	3	35
7	3	4	35
7	2	5	21
6	5	1	6
6	4	2	15
6	3	3	20
6	2	4	15
5	4	1	5
5	3	2	10
5	2	3	10
4	3	1	4
4	2	2	6
3	2	1	3

These combinations of datasets (Cr) are then used as a criterion to search for common genes across r number of gene expression profiles when comparing n number of datasets fail to yield any common genes. However in this scenario, n number of datasets are still used as the search space from which all possible combinations (Cr) of r datasets are compared to each other in order to increase the probability of finding a common gene. Once common genes have been identified using this method, MicroPath will report the results to the interface.

2.3. Extracting Gene Expression Patterns Graphically and Statistically

Following the identification of common genes across n datasets using either of the methods described earlier, the next stage in the analysis is to generate a graphical representation of this expression data from which biologically meaningful patterns can be extracted. Because signals pertaining to transcriptome states tend to be diluted over entire profiles, a specific criterion is required to narrow down the common genes of interest to include only those genes that are consistently regulated according to the biological question. The assumption we have made is that any

given common gene across n datasets can exhibit one of three specific behaviours. It can either be consistently upregulated across all datasets, downregulated across all datasets and up or downregulated across all datasets. Based on the nature of the specific biological question, users can select the appropriate pattern from the options, which will result in a graphical display of those genes which satisfy the search criteria. Together with this faculty to graphically extract patterns for individual gene expression data points, MicroPath also implements the pearsons correlation coefficient statistical test in order to extract a global gene expression pattern existing between common genes pertaining to two individual expression profiles. The correlations are calculated in a pair-wise manner until each expression data has been statistically compared to all other datasets within n , according to the pearsons correlation coefficient equation:

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}}$$

Each pair-wise score is then finally averaged in order to provide a global measure of correlation existing between n expression profiles. Scores are reported from -1 (perfect negative correlation) to +1 (perfect positive correlation).

2.4. High Throughput Molecular Pathway Analysis

To decipher molecular mechanisms fundamental to the researcher's biological question, it is necessary to map common gene expression profiles of co-expressed genes to molecular pathways. This is because biological pathways reveal molecular dependencies that exist between genes by illustrating how they collaborate with one another when they participate in specific biological functions. Furthermore, pathways reveal various signalling cascades that play imperative roles in dictating these gene associations. In light of this, we have implemented MicroPath to access the Application Programming Interface (API) of the molecular pathway database belonging to KEGG [13] using SOAP-Lite in order to dynamically interact with the static pathway maps. Perl scripts were

written for MicroPath to specifically 1) search for user's co-expressed genes in all biological pathways, 2) highlight genes on to pathways, and 3) return the results of the search to MicroPath's interface (i.e. URL's of colour coded pathway maps) (Figure 2). Once MicroPath has searched for all of the user's co-expressed genes in all of the molecular pathways, the URL of each pathway is displayed on the sub-interface. In order to avoid redundancy issues, the URL for each pathway will highlight all co-expressed genes that participate in a given pathway. To help users identify biologically meaningful pathways relevant to their specific biological question, MicroPath will calculate the number of genes identified in a given pathway and 1) express this as a percentage in relation to the total number of common genes from the intersection and 2) express this as a percentage in relation to the total number of genes belonging to that pathway.

Clicking on these links will generate the specific KEGG pathway in HTML on which users co-expressed genes will be highlighted.

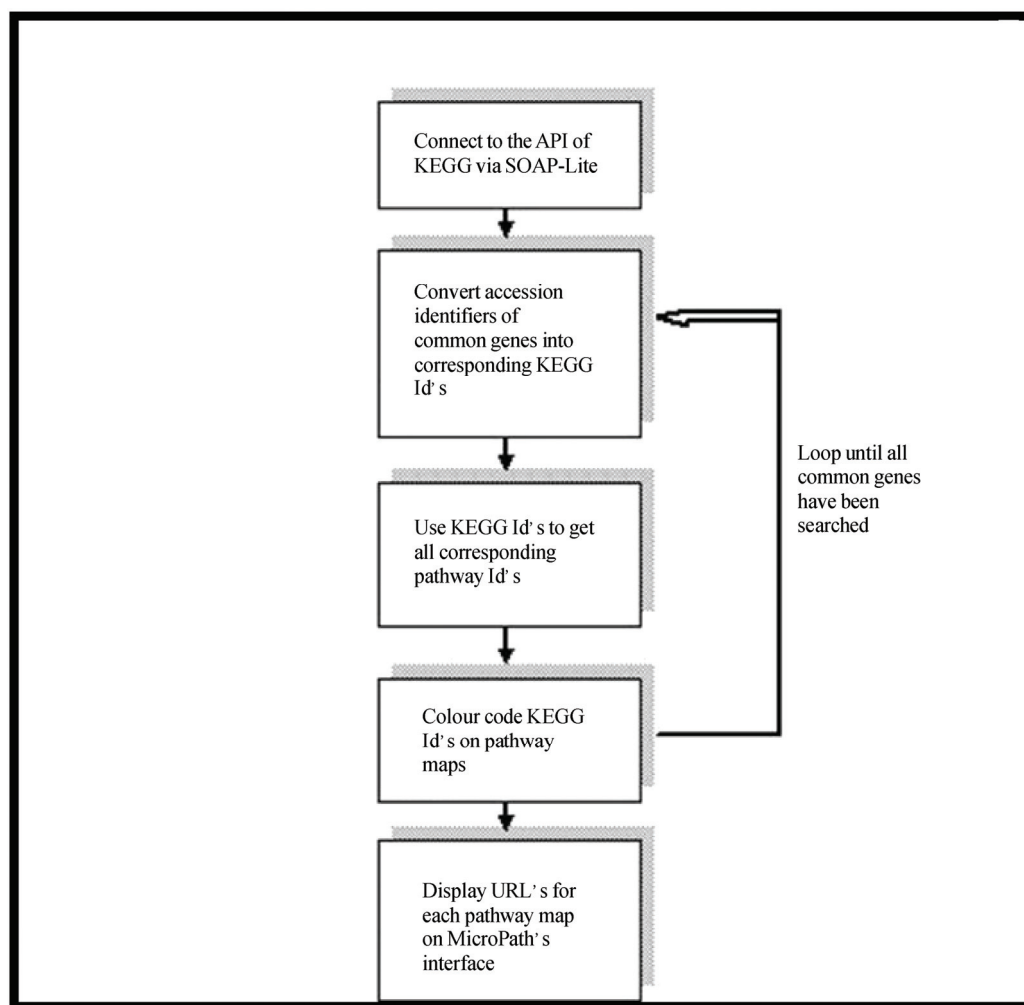


Figure 2. Flow diagram of how MicroPath carries out high throughput molecular pathway analysis by connecting to the API of KEGG.

2.5. Generating and Processing Gene Expression Datasets

Gene expression datasets used for the purpose of this article were generated from our in-house microarray experiments as well as published datasets, where the fold change approach was used to select a set of differentially expressed genes from pre-processed data. Matchminer [14] and the Synergizer [15] tools were used to convert gene Hugo identifiers and long names into Genbank accession Id's in order to ensure that the gene identifiers were of the same type across all datasets prior to comparison. Raw expression data was generated, filtered and normalised using GenePix pro 4.1 [16] and Acuity 4.0 [17] software. Although we used cDNA microarray data for the purpose of demonstrating MicroPath's capabilities, other data types generated from different platforms such as affymetrix can also be analysed provided Genbank accession identifiers are used to represent the genes.

3. RESULTS AND DISCUSSION

Regardless of the biological question, a typical microarray experiment almost always results in the generation of a set of differentially expressed genes, which represents genes of most importance to the biologist. Therefore, by carrying out several biologically related microarray experiments, several sets of differentially expressed genes would be generated, which would need to be compared and mined efficiently in order to help answer the biological questions asked by the investigators from different research laboratories around the world. Employing manual methods of comparison in this situation would be very inefficient and infeasible. In light of this, to demonstrate the benefits that can be derived from analysing multiple gene expression profiles using MicroPath, we employed datasets generated from our in-house microarray experiments as well as published data. The biological question related to these studies focussed on unravelling the underlying molecular mechanisms dictating immune tolerance by analysing the role of Egr-2 in implicating T-cell tolerance. Although the Early Growth Response gene (Egr-2) has been recently characterised as a candidate tolerance-inducing transcription factor, which interacts with specific genes in order to induce the state of T-cell tolerance [18,19], the possibility of further putative unknown target genes exists that may be vital to the mechanism of tolerance. Hence, the biological purpose of our experiments was to attempt to identify such potentially important genes via the comparison of biologically related expression datasets using MicroPath.

Data consisting of a set of differentially expressed genes generated from the comparison of tolerance Vs activated mice CD4+ T cells was obtained from the ArrayExpress website (accession number: e-mexp-283). The first in-house experiment aimed to generate differentially expressed genes from the comparison of an un-stimulated T cell line from which the Egr-2 gene had been knocked out and a wild type un-stimulated cell line.

The second in-house experiment focussed on the comparison between an Egr-2 knock-out T cell line activated with CD3/CD28 for 6 hours and a wild type cell line also activated with CD3/CD28 for 6 hours. Results generated from these experiments were then compared with the aforementioned published tolerance data using MiNER in order to understand the molecular mechanisms controlling immune tolerance.

3.1. Comparison of Gene Expression Profiles Pertaining to Immune Tolerance

The first step in the analysis was to subject the above-mentioned expression profiles to MicroPath in order to identify genes amongst them that had the same accession identifiers. Having done this, MicroPath identified 31 differentially expressed genes that were common to all three expression datasets and generated a graph to delineate their expression values (**Table 2, Figure 3**). A simple number crunching exercise was used to perform this task since its use generated a reasonable number of common genes, which did not warrant the use of permutations and combinations to perform the search. The next step was to use these 31 differentially expressed genes as a search space to determine those genes that have the potential to be co-expressed. In order to do this, we employed MicroPath's graphical utility to extract gene expression patterns, which led to the identification of 6/31 genes that were found to be upregulated in tolerance Vs activated CD4+T-cells and downregulated in both p-KOA0 Vs WTA0 and p-KOA6 Vs WTA6 datasets (**Table 2**). The remaining 25 common differentially expressed genes were found to be highly and lowly expressed in tolerance and knock-out datasets respectively. Statistical analysis revealed an overall pearson's correlation score of 0.109 from the pair-wise comparison of tolerance data with p-KOA0 Vs WTA0 and a score of -0.123 from the comparison of tolerance with p-KOA6 Vs WTA6. Furthermore, Reverse Transcriptase PCR experiments confirmed that 15 genes from our tolerance Vs activated data were found to be highly expressed in immune tolerance and from these 15 genes, 8 were found to be common amongst all three expression profiles (**Table 2**).

Because Egr-2 has been previously characterised and found to be highly upregulated in immune tolerance, these results generated from MicroPath are biologically significant because as expected, those genes that were highly expressed in our tolerance Vs activated datasets were found to be insignificantly expressed in our p-KOA6 Vs WTA6 and p-KOA0 Vs WTA0 datasets (from which the Egr-2 gene was knocked out of the cell lines). Amongst these genes, Ap1s1, Shd, Surf6, Vil2, Lilrb4, Tbx21 and Pcd11g2 (**Table 2**) have been confirmed to be upregulated in the process of immune tolerance [20], all of which were found to exhibit low expression values in our knock-out expression datasets. This consistent gene expression pattern can be seen graphically in **Figure 3**. However, from the 31 interesting common genes, 16 were not confirmed to be involved in

Table 2. Tabulated overview of gene accession ids, Hugo ids and fold change values belonging to 31 common genes identified from the comparison of tolerant Vs activated CD4+ T cells, p-KOA0 Vs WTA0 and p-KOA6 Vs WTA6 expression datasets. Entries highlighted in bold represent genes that were found to be up-regulated in tolerance Vs activated CD4+ T cells and down-regulated in both p-KOA0 Vs WTA0 and p-KOA6 Vs WTA6 datasets. Entries with * represent genes that have been confirmed to be highly expressed in tolerance by RT-PCR.

Gene ID	HUGO ID	Fold Change (p-KOA0 Vs WTA0)	Fold Change (p-KOA6 Vs WTA6)	Fold Change (Tolerance Vs activated)
NM_007381	Acadl	0.371336	0.624525	6.373
NM_007457	Ap1s1 *	0.542474	0.31525	4.965
NM_007664	Cdh2	0.243646	-0.7999	1.658
NM_008205	H2-M9	-0.08048	0.116434	2.857
NM_008972	Ptma	-1.31334	-0.46688	5.42
NM_009128	Scd2	-0.18816	-0.39366	4.552
NM_009168	Shd *	-0.17495	-0.53582	2.838
NM_009298	Surf6 *	0.272072	0.126301	4.365
NM_009465	Axl	0.149539	1.475806	3.836
NM_009510	Vil2 *	-0.49824	0.319645	3.151
NM_010102	Edg6	0.313489	0.132689	1.573
NM_010413	Hdac6	-0.90335	-0.8226	4.745
NM_010548	Il10 *	3.083863	1.660739	3.521
NM_010638	Bteb1	0.024803	-0.42533	1.613
NM_011125	Pltp	-0.5354	-0.71558	4.363
NM_011620	Tnnt3	-0.61646	0.035844	1.665
NM_011696	Vdac3	-0.98084	0.191964	4.701
NM_011705	Vrk1	0.466922	-0.34601	2.032
NM_013488	Cd4	0.584494	0.420277	4.905
NM_013490	Chka	-2.13728	-0.69458	5.677
NM_013532	Lilrb4 *	0.792335	1.110898	2.111
NM_013615	Odf2	2.776384	3.004449	4.809
NM_013814	Galnt1	-0.47752	0.500297	2.246
NM_013866	Zfp385	0.118995	0.428591	1.664
NM_016772	Ech1	-0.0666	0.053081	4.284
NM_019507	Tbx21 *	0.124767	-0.32731	1.595
NM_019561	Ensa	0.778767	-0.44703	1.718
NM_019777	Ikbke	0.291602	-0.00772	1.609
NM_020027	Bat2	0.291219	-0.23966	5.091
NM_021396	Pdcd1lg2 *	1.140087	0.079182	3.921
NM_021538	Cope	0.154049	0.264541	2.035

tolerance by RT-PCR yet some of them also exhibited a coherent pattern of gene expression. For example, Ptma, Scd2, Hdac6, Pltp and Chka were all highly expressed in tolerance and conversely downregulated in both knock out datasets. There is a possibility that these genes may also be insignificantly expressed due to the absence of Egr-2. However, conducting RT-PCR for these specific genes would be required in order to confirm that their over-expression results in T-cell tolerance.

3.2. Deciphering Gene Regulatory Networks of Co-Expressed Genes Via High Throughput Molecular Pathway Analysis

The final stage of the analysis entails using MicroPath's function to connect to the Application Programming Interface (API) of KEGG via SOAP-Lite in order to carry out high throughput molecular pathway analysis. Therefore, for this stage in the analysis, we used MicroPath to map 31 of our co-expressed interesting genes to KEGG pathways and from these 31 genes, 14/31 were identified in a total of 31 molecular pathways (Table 3). Interestingly, several of these pathways were related to the study of immunology and illustrated biological networks such as MapKinase, Jak-Stat, T-cell receptor signalling and

Cytokine-cytokine interactions. More specifically, the Pdcd1lg2 gene (accession id: NM_021396) was identified in the Cell Adhesion Molecules (CAM) pathway (Table 3) and studies have confirmed that the over-expression of Pdcd1lg2 has resulted in consistently low levels of Interleukin-2 (IL-2) in naive CD4(+) T-cells [21]. Further studies have correlated the over-expression of this gene to the negative regulation of T-cell activation. In one particular study, PDL2 (Pdcd1lg2) deficient mice were created in order to characterise the function of this gene in T-cell activation and tolerance, and results generated from this study suggested that Antigen-presenting cells from PDL2-deficient mice were found to be more potent in activating T-cells in vitro when compared to the wild-type counterparts [22]. These findings are conclusive and correlate well with the results generated from our in-house microarray experiments because using MicroPath to compare all three of our datasets followed by extracting gene expression patterns from them resulted in an important finding that Pdcd1lg2 was not only found to be over-expressed in tolerance (fold change of 3.921), but it was also under-expressed in our KOA0 Vs WTA0 and KOA6 Vs WTA6 knock-out datasets (with a fold change of 1.140 and 0.079 respectively) (Table 2). This

particular finding is in agreement with the aforementioned studies, concluding that Pcd11g2 has a negative inhibitory role towards the process of T-cell activation. In addition, molecular pathway analysis of the Interleukin-10 (IL-10) gene using MicroPath, identified its role in the Cytokine-cytokine interaction, Jak-STAT and T-cell receptor signalling pathways; all three of which are important immunological pathways. IL-10 is a well known cytokine, which has previously been shown to successfully induce immune tolerance in Dendritic Cells [23]. Results generated from MicroPath revealed that IL-10 was highly expressed in our tolerance data with a fold change of 3.521, which was found to be expressed lower in our KOA0 Vs WTA0 profile (fold change: 3.084). Interestingly, following activated with

CD3/CD28 for 6 hours, its expression dropped significantly to 1.66, perhaps attributable to the absence of Egr-2. Likewise, other genes from the 31 co-expressed interesting genes show similar patterns of expression and perhaps may be candidate genes for Egr-2 mediated T-cell tolerance. However, this is yet to be confirmed by publications. Finally, the pathway analysis function of MicroPath was used to calculate the percentage of genes identified in each pathway in relation to 1) the intersection of common genes and 2) the total number of genes comprising each pathway. From the results, the Cell Adhesion Molecules (CAM) pathway was particularly significant since 12.91% of the overall pathway was affected by 6.84% of genes common to all 3 expression profiles (Table 4).

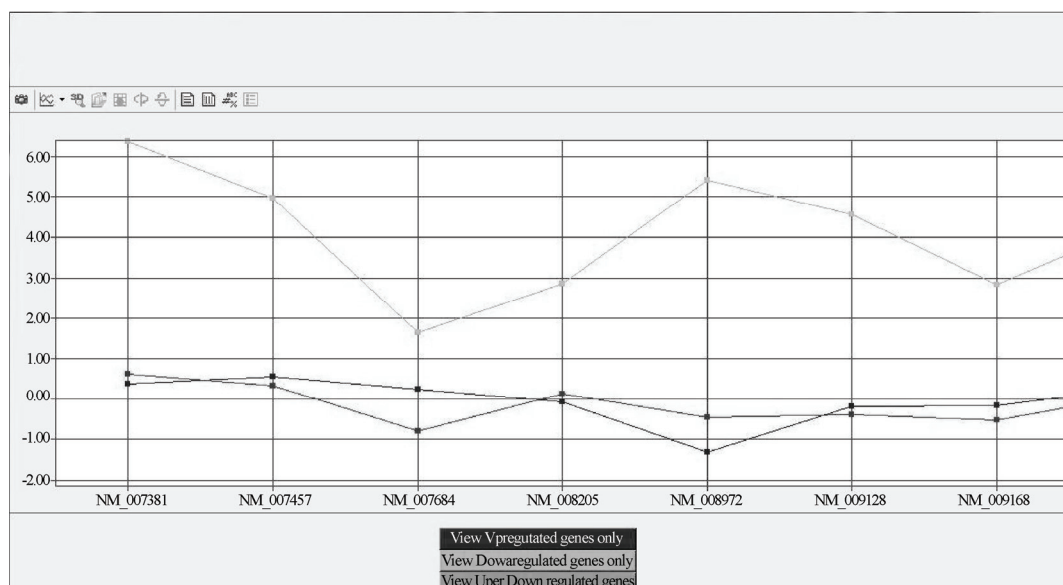


Figure 3. A preliminary graphical overview of common interesting genes generated from the comparison of tolerant Vs activated CD4+ T cells (green), p-KOA0 Vs WTA0 (red) and p-KOA6 Vs WTA6 (blue) expression datasets. It can be seen that genes that are highly expressed in tolerance appear to be expressed poorly in the knock-out datasets. This pattern is consistent throughout the 31 gene expression data points.

Table 3. Tabulated data generated from high throughput molecular pathway analysis of co-regulated genes. 14/31 common interesting genes were identified in a total of 31 molecular pathway maps of KEGG.

GenBank Accession ID	HUGO ID	Pathway ID	Total No of pathways	GenBank Accession ID	HUGO ID	Pathway ID	Total No of pathways		
NM_007381	Acadl	mmu00071	5	NM_009510	Vil2	mmu04670	2		
		mmu00280				mmu04810			
		mmu00410							
		mmu00640							
		mmu03320							
NM_007664	Cdh2	mmu04514	1	NM_008205	H2-M9	mmu04514	3		
NM_013488	Cd4	mmu04612	4			NM_013814		Galnt1	mmu00512
		mmu04640							mmu01030
mmu04660									
NM_011696	Vdac3	mmu04020	1	NM_019777	Ikake	mmu04010	2		
NM_011125	Pltp	mmu03320	1	NM_010102	Edg6	mmu04620	1		
NM_016772	Ech1	mmu00350	3	NM_021396	Pcd11g2	mmu04080	1		
		mmu00362				mmu04514			
		mmu00628							
		mmu04060							
NM_010548	Il10	mmu04630	3	NM_013652	Ccl4	mmu04060	2		
		mmu04660				mmu04620			

The fundamental strength of MicroPath stems from the implementation of a novel search strategy for the comparison of multiple gene expression profiles. Although there are a few software that cater for multiple gene expression comparison, there is currently no software that searches for common genes beyond simple number crunching methods of comparison (Table 5). Just because a direct comparison of a given number of datasets may not yield any common genes, it

does not mean that the analysis should end here since there is a potential to identify common genes across $n-1$ profiles. MicroPath ensures that such genes are identified, which current software would overlook. When coupled with other important functions such as pattern extraction and pathway analysis, it becomes apparent that MicroPath would offer valuable assistance to biologists wanting to decipher their high throughput data.

Table 4. Results generated from pathway analysis showing the extent to which each pathway is affected by common genes from the intersection. The percentages reflect the proportion of common genes that contribute towards controlling the proportion of each pathway.

Pathway ID	Pathway Name	GenBank Accession ID	Result from Analysis
mmu00071	Fatty Acid Metabolism	NM_007381	3.26% of genes contribute 8.45% role in pathway
mmu00280	Valine, leucine and isoleucine degradation	NM_007381	3.26% of genes contribute 2.73% role in pathway
mmu00410	Beta Alanine Metabolism	NM_007381	3.26% of genes contribute 7.14% role in pathway
mmu00640	Propanoate Metabolism	NM_007381	3.26% of genes contribute 5.88% role in pathway
mmu03320	PPAR Signalling Pathway	NM_007381	3.26% of genes contribute 1.92% role in pathway
mmu04514	Cell Adhesion Molecules	NM_007664	12.91% of genes contribute 6.84 % role in pathway
		NM_008205	
		NM_013488	
		NM_021396	
mmu04612	Antigen Processing & Presentation	NM_013488	3.26% of genes contribute 2.44% role in pathway
mmu04640	Hematopoietic Cell Lineage	NM_013488	3.26% of genes contribute 0.76 % role in pathway
mmu04660	T Cell Receptor Signalling Pathway	NM_013488	6.45 % of genes contribute 3.33 % role in pathway
		NM_010548	
mmu04020	Calcium Signalling Pathway	NM_011696	3.26% of genes contribute 2.33 % role in pathway
mmu00350	Tyrosine Metabolism	NM_016772	3.26% of genes contribute 2.17 % role in pathway
mmu04060	Cytokine-cytokine receptor interaction	NM_010548	6.45 % of genes contribute 0.73 % role in pathway
		NM_013652	
mmu04630	JAK-STAT Signalling Pathway	NM_010548	3.26% of genes contribute 3.85 % role in pathway
mmu04670	Leukocyte Transendothelial Migration	NM_009510	3.26% of genes contribute 1.25 % role in pathway
mmu04810	Regulation of Actin Cytoskeleton	NM_009510	3.26% of genes contribute 1.47 % role in pathway
mmu04940	Type I Diabetes Mellitus	NM_008205	3.26% of genes contribute 4.35 % role in pathway
mmu00512	O-Glycan Biosynthesis	NM_013814	3.26% of genes contribute 10 % role in pathway
mmu04010	MAPK Signalling Pathway	NM_019777	3.26% of genes contribute 0.83 % role in pathway
mmu04620	Toll-Like Receptor Signalling Pathway	NM_019777	6.45% of genes contribute 1.32 % role in pathway
		NM_013652	
mmu04080	Neuroactive Ligand-Receptor Interaction	NM_010102	3.26% of genes contribute 1.15 % role in pathway

Table 5. Functional comparison of MicroPath to similar software packages and applications.

Function	MicroPath	EXPANDER [24]	INCLUSIVE [25]	Pathway Studio [26]	KEGG [13]	BioCarta [27]	MaXlab [28]
Suitable for high throughput data analysis	YES	YES	YES	YES	NO	NO	YES
Suitable for comparing multiple gene expression profiles	YES	YES	NO	YES	NO	NO	YES
Implementation of efficient algorithm to search for common genes from $n-1$ datasets	YES	NO	NO	NO	NO	NO	NO
Graphical representation of gene expression values from multiple datasets	YES	NO	NO	NO	NO	NO	YES
Pattern extraction from Graph data	YES	NO	NO	NO	NO	NO	NO
Construction of pathway maps	YES	NO	NO	YES	YES	YES	NO
Mapping gene expression data to pathway maps	YES	NO	NO	YES	NO	NO	NO
User interactive software (S) or Database (D)	S	S	S	S	D	D	S

4. Conclusion

In this article, we have illustrated the potential benefits that can be derived from using MicroPath for the analysis of multiple gene expression profiles. Each function of the software has been developed to streamline the overall analysis pipeline, providing users with a walkthrough of how their data is biologically deciphered. Here, we have applied to our software, microarray datasets generated from different laboratories pertaining to the molecular mechanisms underlying immune tolerance. However, MicroPath is capable of analysing data for any given biological question, whether the datasets are taken from public repositories such as ArrayExpress or generated from in-house microarray experiments. We believe that its faculty to use both number crunching and permutations and combinations as the search strategy to identify the intersection of common genes, coupled with its function to extract gene expression patterns graphically and statistically makes this an attractive software for biologists to use. Finally, its ability to carry out live streaming of mapping genes to biological pathways makes it a useful tool for the automation of multiple gene expression analysis.

Availability and requirements

Project name: MicroPath

Project home page: www.1066technologies.co.uk/micropath

Operating system(s): MicroPath has been tested on Windows 2000, XP and Vista

Programming language: Visual Basic.Net, Perl

Other requirements: None

License: N/A

Any restrictions to use by non-academics: No

Acknowledgements

This study was supported by grants from the UK Medical Research Council (MRC) (Grant number: G0300520).

REFERENCES

- [1] U. Sarkans, H. Parkinson, G. G. Lara, A. Oezcimen, A. Sharma, N. Abeygunawardena, S. Contrino, E. Holloway, P. Rocca-Serra, G. Mukherjee, M. Shojatalab, M. Kapushesky, S. A. Sansone, A. Farne, T. Rayner and A. Brazma. (2005) The ArrayExpress gene expression database: a software engineering and implementation perspective. *Bioinformatics* 21(8): 1495–1501.
- [2] T. Barrett and R. Edgar. (2006) Mining Microarray Data at NCBI's Gene Expression Omnibus (GEO). *Methods Mol Biol* 338: 175–190.
- [3] D. Ghosh, Barrette, T. R., Rhodes, D. and Chinnaiyan, A. M. (2003). Statistical issues and methods for meta-analysis of microarray data, A case study in prostate cancer. *Funct. Integr. Genomics* 3, 180–188.
- [4] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh and A. M. Chinnaiyan, (2002). Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.* 62, 4427–4433.
- [5] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey and A. M. Chinnaiyan (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. USA* 101, 9309–9314.
- [6] J. Wang, K. R. Coombes, W. E. Highsmith, M. J. Keating and L. V. Abruzzo (2004). Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: A meta-analysis of three microarray studies. *Bioinformatics* 20, 3166–3178.
- [7] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu and Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome Res.* 17, 1537–1545.
- [8] J. Stelling, (2004). Mathematical models in microbial systems biology. *Curr. Opin. Microbiol.* 7, 513–518.
- [9] G. Joshi-Tope, M. Gillespie, I. Vasrik, P. D'Eustachio, E. Schmidt, B. de Bone, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, *et al.* (2005). A knowledgebase of biological pathways. *Nucleic Acids Res.* 33, D428–D432.
- [10] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.* (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 102: 15545–15550.
- [11] S. Khalid, F. Fraser, M. Khan, P. Wang, X. Liu and S. Li, (2006a). Analysing Microarray Data using the Multi-functional Immune Ontologiser. *J. Integrative Bioinformatics* 3, 25.
- [12] S. Khalid, M. Khan, P. Wang, X. Liu and S. -L. Li, (2006b). Application of bioinformatics in the design of gene expression microarrays. *Second International Symposium on Leveraging Applications of Formal Methods, Verification and Validation (isola 2006)*, pp. 146–160.
- [13] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno and M. Hattori, (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32.
- [14] K. J. Bussey, D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W. C. Reinhold, B. Zeeberg, W. Ajay and J. N. Weinstein, (2003) MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biology.* 4, R27.
- [15] G. F. Berriz and F. P. Roth, The Synergizer service for translating gene, protein, and other biological identifiers. (2008). *Bioinformatics*. [Epub ahead of print].
- [16] GenePix pro 4.1: <http://www.axon.com>
- [17] Acuity 4.0: http://www.moleculardevices.com/pages/software/gn_acuity.html
- [18] M. Safford, S. Collins, M. A. Lutz, A. Allen, C. Huang, J. Kowalski, A. Blackford, M. R. Horton, C. Drake, R. H. Schwartz and J. D. Powell, (2005) Egr-2 and Egr-3 are negative regulators of T cell activation. *Nature Immunology* 6 472–480.
- [19] L. E. Warner, J. Svaren, J. Milbrandt and J. R. Lupski, (1999) Functional consequences of mutations in the early growth response 2 gene (EGR2) correlate with severity of human myelinopathies. *Hum. Mol. Genet.* 8 1245–1251.
- [20] P. O. Anderson, B. A. Manzo, A. Sundstedt, S. Minaee, A. Symonds, S. Khalid, M. E. Rodriguez-Cabezas, K. Nicolson, S. Li, D. C. Wraith and P. Wang, (2006) Persistent antigenic stimulation alters the transcription program in T cells, resulting in antigen-specific tolerance. *European Journal of Immunology.* 36, 1374–85.
- [21] H. Kuipers, F. Muskens, M. Willart, D. Hijdra, F. B. van Assema, A. J. Coyle, H. C. Hoogsteden and B. N. Lambrecht (2006). Contribution of the PD-1 ligands/PD-1 signaling pathway to dendritic cell-mediated CD4 (+) T cell activation. *European Journal of Immunology.* 36 (9), 2472–82.
- [22] Y. Zhang, Y. Chung, C. Bishop, B. Daugherty, H. Chute, P. Holst, C. Kurahara, F. Lott, N. Sun, A. A. Welcher and C. Dong, (2006). Regulation of T cell activation and tolerance by PDL2. *Proc Natl Acad Sci U S A*, 103(31), 11695–11700.

- [23] X. Li, K. Dou, H. Liu, F. Zhang and L. Cai, (2007). Immune tolerance induced by IL-10 and methylprednisolone modified dendritic cells in vitro. *Chinese Journal of cellular and molecular Immunol.* 23 (5), 436-8.
- [24] R. Shamir, A. Maron-Katz, A. Tanay, C. Linhart, I. Steinfeld, R. Sharan, Y. Shiloh and R. Elkon, (2005) EXPANDER-an integrative program suite for microarray data analysis. *BMC Bioinformatics*, 6: 232.
- [25] G. Thijs, Y. Moreau, F. D. Smet, J. Mathys, M. Lescot, S. Rombaux, P. Rouze, B. D. Moor and K. Marchal, (2002) INCLUSive: Integrated Clustering, Upstream sequence retrieval and motif Sampling. *Bioinformatics*, 18, 331-332.
- [26] A. Nikitin, S. Egorov, N. Daraselia and I. Mazo,. (2003) Pathway studio-the analysis and navigation of molecular networks. *Bioinformatics*, 19, 2155-2157.
- [27] BioCarta, Charting pathways of life. <http://www.biocarta.com>.
- [28] S. Khalid, M. Khan, C. B. Gorle, K. Fraser, P. Wang, X. Liu and S. Li, MaXlab: A novel application for the cross comparison and integration of biological signatures from microarray studies. In *Silico Biology* 8, 0029: 2008.