A Pathway-based Gene Selection Method Provides Accurate Disease Classification

Maysson Al-haj Ibrahim[1, 2], Sabah Jassim[1], Michael A Cawthorne[2], and Kenneth Langlands[2]
[1]*Department of Applied Computing, The University of Buckingham*
[2]*The Clore Laboratory, The University of Buckingham, UK*

Abstract

Transcriptional profiling techniques facilitate the identification of biomarkers: genes that provide insight into disease diagnosis and natural history. Biomarker discovery can be modelled as a feature selection problem that aims to find the most discriminating features (i.e. gene expression patterns) for accurate disease classification, particularly in identifying those individuals at risk of more aggressive disease. Typical feature selection algorithms identify individual genes, and the disease discrimination power of each gene is considered separately. We describe a gene selection method incorporating prior biological knowledge of genetic pathways to find groups of strongly-correlated genes that accurately discriminate complex as well as simple disease traits. The first step is to identify the most impacted biological processes in a genome-wide expression (microarray) dataset describing a particular pathological condition. A specified number of differentially-expressed genes from relevant pathways are then selected for disease classification. The advantage of this method is that it searches for a related group of strongly-correlated genes rather than individual ones, with greater relevance to the disease type or clinical grade. We compared our algorithm to existing feature selection and ranking methods using open-access datasets derived from a series of well characterised clinical conditions, and two classification methods: K-Nearest Neighbour (KNN) and Support Vector Machine (SVM). Our algorithm consistently outperformed other methods in terms of disease classification accuracy. Moreover, we were able to reduce the number of genes required to accurately discriminate disease states, critical in the creation of diagnostic tests.

1. Introduction

Alterations in gene expression underlie all disease states, although until recently it was not possible to determine all gene activity in clinical samples. Recent advances in transcriptional profiling technology (microarrays) have made this a reality, but mining complex datasets for relevant insights is one of the major challenges in translational biology. The selection of genes most relevant to disease classification (biomarkers) is a challenging task in high dimension space such as microarrays, where the expression profile of thousands of genes is measured simultaneously. Biomarkers are typically selected by measuring the power of their expression profiles to discriminate between different disease states as well as their ability to provide useful diagnostic, mechanistic and prognostic information [*1*].
Biomarker discovery can be modelled as a feature selection problem that aims to find the most discriminating features (genes) for accurate disease classification.

Feature selection methods can be broadly categorized into filter and wrapper methods [*2*]. In filtering methods, a gene list is ranked according to pre-defined criteria, and the discriminatory power of each gene is considered separately. The most frequently used ranking criteria are: t-test, relative entropy, Bhattacharyya distance, Wilcoxon test and fold-change. Wrapper methods identify a small subset of *r* out of *n* features that minimize the classification error where $r << n$. As *r* may be predefined, a straightforward approach to select the best *r* features out of *n* is to try all possible combinations *C* where:

$$C_r^n = \frac{n!}{r!(n-r)!} \qquad (1)$$

For example, if we want to select the most informative subset of 10 out of 100 genes with a minimum classification error (r=10, and n=100), then 1.731×10^{13} classification attempts are required. However, as genes assayed in a typical microarray experiment number in the tens of thousands, wrapper methods are very computationally-demanding even when using state-of-the-art algorithms such as branch and bound search, sequential forward/backward selection, or sequential forward/backward floating searching [*3*,*4*,*5*]. Although filtering methods are classifier-independent and fast compared to wrapper methods, their main deficiency is that each gene is examined individually and relationships between groups of genes are ignored. As consequence of being classifier-dependent and computationally-expensive wrapper methods are not widely used in biomarker discovery research, unlike filter methods, which have been extensively reported [*2*].

Biologists have gone to great lengths in recent years to classify genes in the context of discrete biological processes (pathways) in order to facilitate the rational analysis of large and complex microarray expression datasets. This has generated a vast repository of biological information, which is curated in publicly-available databases. Classification has taken different forms, including categorizing genes according to narrowly-defined descriptive terms (cellular component, biological process and molecular function) by

the Gene Ontology (GO) consortium [*6*], or by grouping genes that act in concert to effect particular outcomes (cellular pathways), such as in the database maintained by the Kyoto Encyclopedia of Genes and Genomes (KEGG) [*7*]. Recently, methods to effectively interrogate microarray data in the context of pathways have emerged as enriching for coordinated groups of genes is more informative to biologists, compared to considering unsorted lists of genes whose expression changes in a pathological condition. [*8*] A number of sophisticated statistical methods are described in the literature to identify those processes whose constituent members showed a greater number of gene expression changes than would be expected by chance, including Fisher exact and Chi-squared tests [*9*]. Others methods, such as gene set enrichment analysis (GSEA) [*10*] and z-score [*11*] assign each pathway a statistical score representing its contribution to the phenotype under analysis. Tools such as GenMapp [*12*], Gene-Sifter [*13*], and Pathway Miner [*14*] use z-scores in evaluating either GO term or pathway enrichment (reviewed in [*9*]). Until recently, the identification of biomarkers relied on correlating individual genes to specified clinical states or outcomes (for example event-free survival). However, several recent papers describe biomarker discovery methods that exploit pathway information. Guo et al [*15*] used an arithmetic mean and median of all the gene expression values in each category defined by GO to capture the activity of that category, represented as a vector. Rapaport et al [*16*] and Chen et al [*17*] relied on Principal Component Analysis (PCA) to summarize all genes in every pathway in a compact representation. Su et al [*18*] computed the log-likelihood ratio comparing different disease phenotypes based on the expression level of each gene. The activity of a given pathway was inferred by combining the log-likelihood ratios of the constituent genes. Tai et al [*19*] used all genes in a pathway with no transformation. Others, such as Chuang et al [*20*] and Hwang et al [*21*] applied a greedy search algorithm to find a subset(s) of discriminating genes in each pathway. Each subset of genes was then summarized using the mean [*21*] or sum of z-scores [*20*]. Although these approaches have demonstrated classification accuracies that are comparable to conventional gene selection algorithms, they are subject to the following limitations: 1) treating all pre-defined pathways or functional categories (GO terms) equally, regardless of their enrichment levels, results in the inclusion of pathways not relevant to the phenotype; 2) these methods treat all genes in a pathway equally, irrespective of gene expression levels and pathway topology; and 3) although summarizing a set of genes by one or more values such as mean, median, sum of log-likelihood ratio, or first principal component(s) of PCA might lead to satisfactory disease classification accuracy, they do not facilitate the identification of those genes germane to our understanding of disease pathogenesis.

Herein, we describe a gene selection method incorporating prior biological knowledge of biological pathways in order to identify groups of strongly-correlated genes able to accurately discriminate complex as well as simple pathological states. In addition to using static predefined pathways knowledge, our method is adaptive in the sense that it involves a pathways ranking process to identify the most relevant pathways perturbed in a given pathological state.

2. Methods

2.1. Enriching biological pathways by z-score

An overview of our method is shown in Figure 1. A total of 108 signalling pathways, whose description files are written in KEGG mark-up language (KGML), were imported from KEGG into Matlab 7.10.0 in a logical structure. Raw microarray data were preprocessed using Robust Multiarray Averaging (RMA) as standard, and a simple fold-change value for each gene calculated. Every transcript was then tested for significance as follows:

$$Fc(G_i) = \begin{cases} abs(Fc_i) & (abs(Fc_i) >= h)\ \&\ (p < p_1) \\ 1 & Otherwise \end{cases} \quad (2)$$

Differentially-expressed genes are assigned absolute fold-change values upon achieving two criteria: the fold change must exceed a specified threshold *h* (we used 1.5 fold for all datasets in this study), and the p-value yielded by comparing expression values from replicated datasets using a simple t-test must be below a threshold p_1 (we used 0.05 throughout). Genes that may be active, but do not meet fold change or t-test criteria are assigned the value 1. A matrix of all expressed microarray genes values was mapped on to pathway structures before the top z-scoring pathways were identified.

The z-score is a standard statistical test under the hypergeometric distribution. It is used as a measure of relative pathway perturbation after superimposing microarray gene expression data on a curated pathway structure. The z-score value was calculated simply by subtracting the expected number of genes meeting the defined criteria from the total number of genes in the pathway, then dividing by the standard deviation of the total number of genes as follows:

$$zscore = \frac{(r - n\frac{R}{N})}{\sqrt{n(\frac{R}{N})(1-\frac{R}{N})(1-\frac{n-1}{N-1})}} \quad (3)$$

Where:

- N is the total number of expressed genes in the dataset.
- R is the total number of significant genes meeting fold-change and t-test criteria.

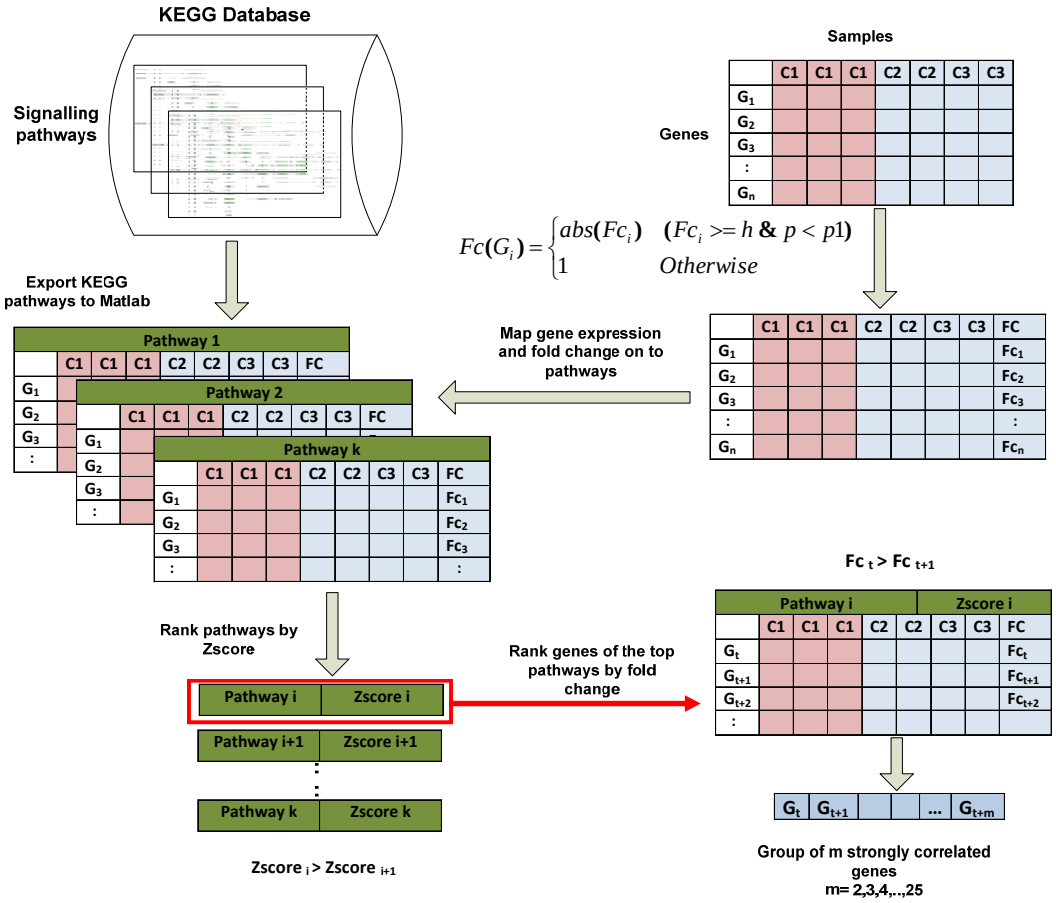

Figure 1. Flow chart outlining the procedure for selecting a vector of strongly correlated genes for disease classification.

- n is the total number of expressed genes in a pathway p_i.
- r is the number of significant genes meeting the criteria (fc>= threshold, p-value < threshold) in the pathway p_i

A positive value indicates that the pathway is likely to contribute to the phenotype as the number of constituent genes whose expression level changes is greater than that expected by a chance. After ranking pathways by z-score, all expressed genes contained within the top *n* pathways were combined in a group. Genes were sorted in descending order by fold change, and the first *m* genes where *m*= {2, 3,.., 25} were selected for disease classification.

2.2. Pathway-based principal component analysis (PCA)

PCA is used to reduce the number of variables in a space by transforming them into another space. We used this method to reduce the number of candidate genes selected from the top *n* pathways ranked by z-score into a smaller number of discriminating meta-genes. A built-in function in MatLab 7.9.0 was used for this purpose.

2.3. Evaluation of classification accuracy

Cross-validation was used to calculate classification accuracy (measured as % error rate, calculated from the number of misclassified samples) by repeatedly partitioning a given microarray data set into two subsets: the training subset and the test subset. K-fold cross-validation (k=10 in our experiments) is selected to partition datasets into k subsets. Training is performed on k-1 subsets and testing on the remaining subset. This process is repeated 10-times as each subset is taken to be a test set in turn (a leave one out method). Two of the most frequently used classifiers, K Nearest Neighbour [*22*] (KNN with K=1, i.e. 1-NN) and Support Vector Machine (SVM) [*23*] were selected for our experiments and applied using built-in functions in MatLab 7.9.0. These two classifiers perform in two distinct ways: KNN measures the distance (in this case the Euclidean distance) between a sample and all training samples to assign it the class of the closest neighbour. SVM uses the training samples to construct an N-dimensional hyperplane that separates the data into two categories, before assigning the test sample to one of these categories.

We tested three conditions. In the first, our Pathway-based Differentially Expressed Genes (PDEGs) method was

compared to five well-known feature ranking and selection algorithms: t-test, relative entropy (also known as Kullback-Leibler distance or divergence), Bhattacharyya distance, Wilcoxon test, and fold-change (see [*2*] for a review of these approaches). The top *m* genes were selected for classification where m= {2, 3,.., 25}. In the second group of experiments, we considered classification accuracy when using genes identified from different pathways, either singly or in combination. Finally, we compared our PDEGs method to another pathway-based gene selection method (PCA).

2.4. Datasets

Acute Myeloid Leukaemia

Acute myeloid leukaemia (AML) is a malignancy of blood cells of the myelocytic lineage. Microarray studies have been used to identify gene expression changes that characterise AML cells in order to identify causative mechanisms and novel therapeutic targets. We re-analysed an Affymetrix GeneChip U133A dataset describing a comparison between normal haematopoietic cells isolated from 38 healthy donors and leukemic blasts from 26 AML patients [*24*] (GEO accession GSE9476 [*25*]).

Psoriatic lesions

Psoriasis is a common skin condition arising from inappropriate keratinocyte activation leading to thickened, inflamed skin. Mechanisms of pathogenesis and the influence of potential immune mediators were investigated in a report using Affymetrix® whole genome U133 plus v2.0 arrays (GEO accession GSE14905, [*25*]). Gene expression was profiled in 28 paired samples of lesional and non-lesional skin from psoriatic patients [*26*].

Glioma

Glioma is a tumour arising from glial cells and a number of histological grades are described. We analysed Affymetrix Human Genome U133B Array data derived from 85 tumours: 26 grade III and 59 grade IV [*27*](GEO accession GSE4412 [*25*]).

3. Results and Discussion

3.1. PDEGs vs. other feature selection methods

For each of the three datasets, z-scoring produced a list of top-down ranked pathways. Tables 1, 2 and 3 present the top five impacted pathways in AML, psoriasis and glioma respectively. In these tables, *r* represents the number of significant (i.e. changing) genes in each pathway and *n* is the total number of expressed genes in each pathway. Figure 2 compares classification accuracies using genes derived from critical pathways to other feature selection methods.

TABLE 1. PATHWAYS IN ACUTE MYELOID LEUKEMIA

rank	*pathway*	*z-score*	*r*	*n*
1	Osteoclast differentiation	4.6725	31	126
2	Antigen processing and presentation	4.0928	20	76
3	Natural killer cell mediated cytotoxicity	4.0617	29	127
4	Acute myeloid leukemia	3.8778	16	58
5	T-cell receptor signalling pathway	3.6974	24	105

Figure 2A and Figure 2B compare classification accuracies based on KNN and SVM methods applied to the AML dataset, using between two and 25 genes. The method described herein outperformed the other selection methods. Moreover, this was achieved using a small number of strongly-correlated genes. For example, using just two genes (FHL2 and FCGR3A), our algorithm achieved a 3.1% classification error rate with both KNN and SVM methods, whereas the lowest error rates achieved by the other five methods using two genes only were 12.5% and 14% based on KNN and SVM respectively.

Table 2 shows the top five scoring pathways in psoriasis and Figure 2C and Figure 2D compare classification accuracies achieved using KNN and SVM respectively. The two charts again show that our algorithm achieves better classification accuracy when compared to the five other feature selection methods. Notably, our method achieves perfect classification accuracy (i.e. a 0% error rate) using 15 and 11 genes based on KNN and SVM classifiers respectively. Furthermore, using two genes only (STAT1 and PPIF), the algorithm achieves a 3.5% classification error rate with either classifier, whereas the lowest error rates achieved by previous methods using two genes only are 8.9% and 7.1% using KNN and SVM respectively.

TABLE 2. PATHWAYS IN PSORIASIS

rank	*pathway*	*z-score*	*r*	*n*
1	Toxoplasmosis	7.7324	69	130
2	Staphylococcus aureus infection	6.8527	35	55
3	Cell adhesion molecules (CAMs)	6.7594	65	132
4	Pathways in cancer	5.9077	124	326
5	Chemokine signalling pathway	5.6387	78	187

Table 3 shows the top five glioma pathways ranked by z-score, and Figure 2E and Figure 2F compare classification accuracies as before. Once more, we achieved superior performance in comparison to the five feature selection methods. Using two genes only (SOCS3 and TGFB2), our algorithm achieves 14.1% and 10.5% classification error rates based on the two classifiers, whereas the lowest error rate achieved by the other five methods using two genes is 22.3%. Glioma represents a more complex classification problem compared to the previous two datasets, in which healthy and pathological states were compared. However, PDEGs method consistently outperforms existing methods.

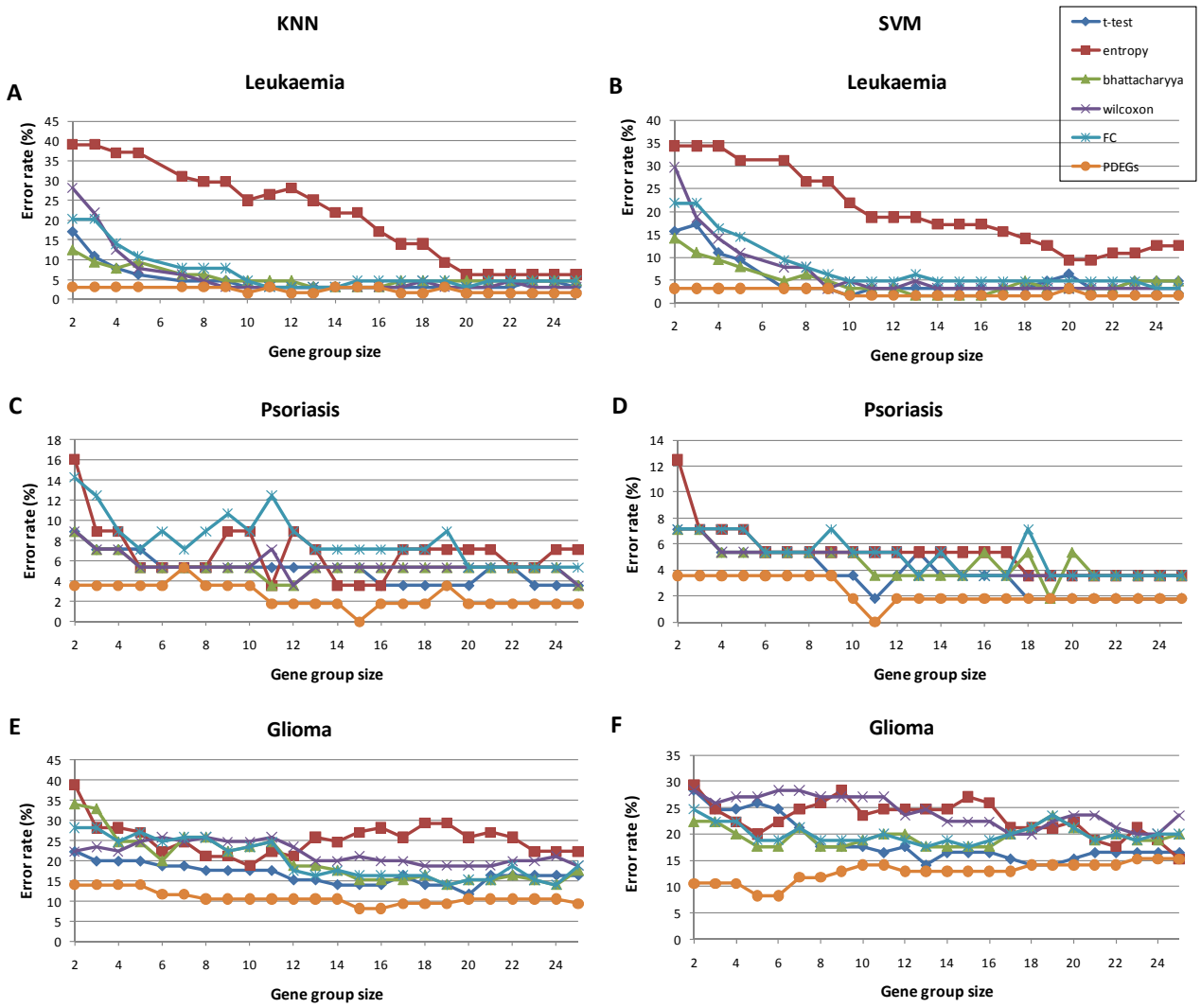

Figure 2. Comparison of PDEGs method to feature-selection methods to classify pathological states. Our pathway-derived method was compared to t-test, relative entropy, Bhattacharyya distance, Wilcoxon test, and Fold Change methods using KNN or SVM classifiers and three datasets: AML, psoriasis and glioma.

TABLE 3. PATHWAYS IN GLIOMA

rank	*pathway*	*z-score*	*r*	*n*
1	Osteoclast differentiation	5.2428	7	34
2	Neuroactive ligand-receptor interaction	3.8116	8	63
3	Chagas disease (American trypanosomiasis)	3.7816	5	30
4	Adipocytokine signaling pathway	3.7424	4	21
5	TGF-beta signaling pathway	3.6851	5	31

3.2. Improved accuracy by combining pathways

We went on to investigate the effect of combining several pathways towards identifying a more informative set of disease-correlated genes. The top five pathways ranked by z-scores were pooled and typical results of glioma classification are shown in Figure 3A (using the KNN classifier) and Figure 3B (using the SVM classifier).

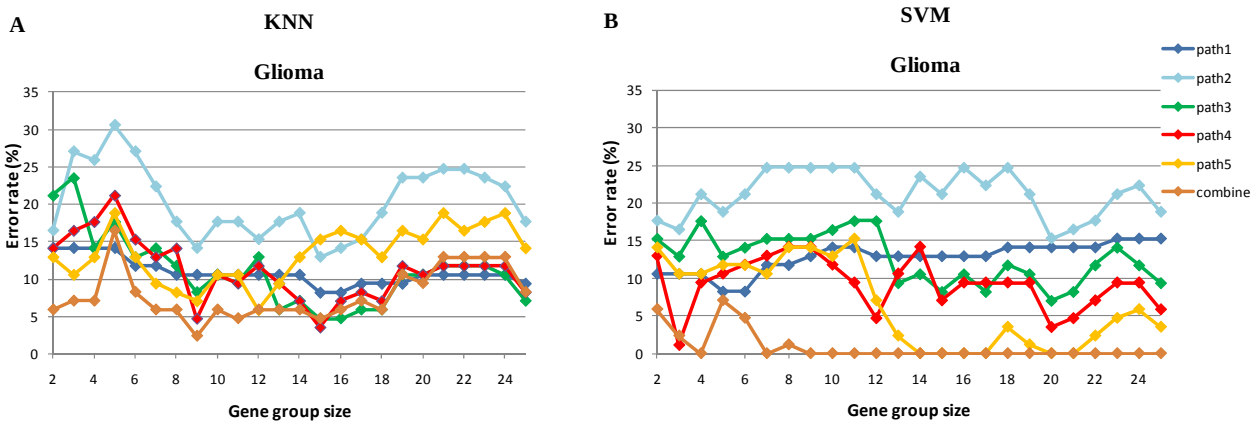

Figure 3. Comparison of multiple pathways combination to the individual pathways using the KNN (A) and the SVM (B) classifiers to study glioma

Figure 3 reveals a number of interesting observations. It is clear that combining pathways provides higher accuracy compared to single pathways. More importantly, even in the challenging glioma dataset, our method was able to achieve zero classification error based on SVM, whereas genes derived from the most informative single pathway achieved minimally a 9% error rate.

3.3. Comparison of pathway-based gene selection methods

We compared our PDEGs method to PCA. In both cases, gene lists were populated by all expressed genes in the top five pathways. In the PDEGs method, genes were ranked by fold-change and the top *m* genes were used for classification. Alternatively, the list of genes was transformed to *m* meta-genes by PCA before classification. In both methods, a different number of discriminating genes m={2,4,…,24} was selected for disease classification. Herein, we use the term meta-genes to emphasise the fact that PCA applies a transformation to the original genes. Critically, the DEGs method has an advantage over PCA in as much as it outputs a group of identifiable genes rather than an abstracted group of metagenes.

Figure 5 compares the two methods using all the three sample datasets, based on KNN and SVM classifiers. Figure 5A and Figure 5B illustrates the superiority of PCA over the DEGs method in discriminating normal from AML blasts when using fewer than 18 genes with either classifier. Similarly, in the psoriasis dataset, PCA has a lower error rate when fewer than 11 genes were selected using KNN and 13 genes using SVM (Figure 5C and Figure 5D). Using a larger group of genes, the classification error rate increases when PCA is used and decreases when the DEG-based method is deployed. In the third dataset, grade III and grade IV glioma, it can be seen from Figure 5E and Figure 5F that PCA more effectively discriminates grade III from grade IV glioma compared to the DEGs method using KNN. Performance is more comparable with SVM, particularly when the number of genes exceeds 12.

4. Conclusions

Traditionally, number of clinical indicators (for example patient age, white blood cell count etc) are used when a new case presents to provide the physician with insight into the possible natural history of the disease. Although this information may be important in stratifying therapeutic protocols, such indicators are limited at predicting, for example, the risk of metastatic disease in those with cancer. Molecular medicine has long promised the early identification of those individuals at risk of an aggressive disease course, and herein we describe a method to more effectively isolate clinically-relevant information from genetic data. We compared the ability of a novel gene selection method (using either pathway-based DEGs or PCA) to classify disease states in comparison to existing methods. We exploited prior knowledge of biological relationships to rank cellular processes perturbed in a microarray dataset in order to create a minimal set of discriminating genes. The DEGs method uses fold-change and t-test criteria to identify the most effective genes whose expression discriminates various conditions. PCA reduces the dimension of the gene candidates' space to generate a more informative list of discriminating meta-genes. This method is powerful when whole genome data is available, although the identification of a small subset of informative genes facilitates the development of relatively inexpensive diagnostic assays. Although both approaches provide increased discriminatory power compared to existing methods, we are in the process of developing DEG-based methods to isolate a minimal gene set able to provide cost-effective and accurate cancer staging at diagnosis. This will ultimately lead to improved disease outcomes by informing the management of disease and facilitating individualised treatment.

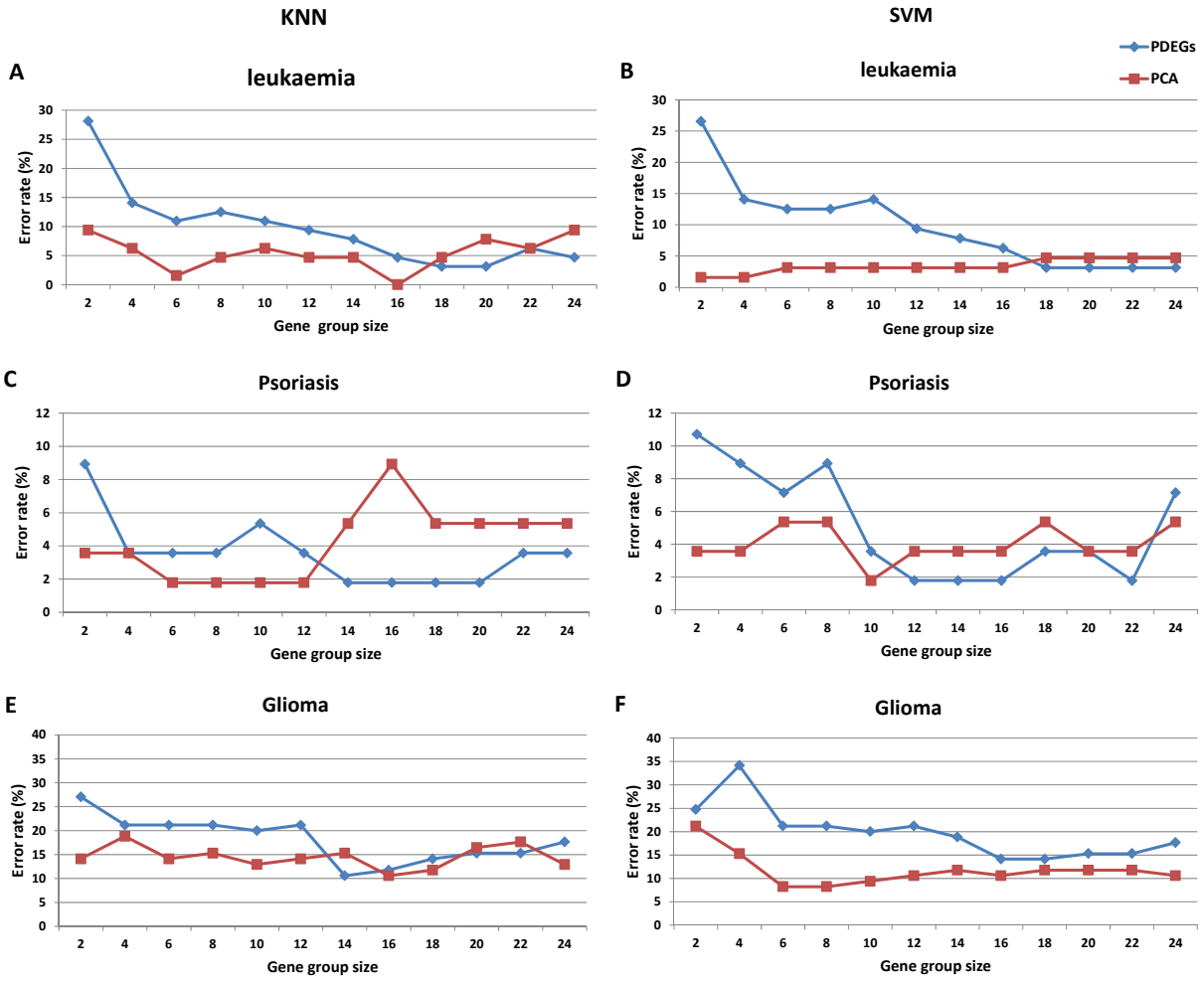

Figure 4. Comparison of PCA and DEG pathway-based gene selection methods using two KNN and SVM different classifiers and three datasets: AML, psoriasis and glioma.

5. References

[1] E. Lee, H.Y. Chuang, J.W. Kim, T. Ideker,and D. Lee., "Inferring Pathway Activity toward Precise Disease Classification," *PLoS computational biology*, vol. 4, p. e1000217, 2008.

[2] M. H. Asyali, D. Colak, O. Demirkaya, and M. S. Inan., "Gene Expression Profile Classification: A Review," *Current Bioinformatics*, pp. 55-73, 2006.

[3] A. K. Jain, R. P. Duin, and J. Mao., "Statistical Pattern Recognition: A Review," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE PAMI*, vol. 22, pp. 4-37, 2000.

[4] A. Jain and D.Zongker., "Feature Selection: Evaluation, Application, and Small Sample Performance," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE PAMI*, vol. 19, pp. 153-157, 1997.

[5] R. Simon., "Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data," *BRITISH JOURNAL OF CANCER*, vol. 89, pp. 1599-1604, 2003.

[6] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, and others, "Gene Ontology: tool for the unification of biology," *Nature genetics*, vol. 25, pp. 25-29, 2000.

[7] The Kyoto Encyclopaedia of Genes and Genomes database. [Online]. http://www.genome.jp/kegg/

[8] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane,and P. J. Park., "Discovering statistically significant pathways in expression profiling studies," *PROCEEDINGS-NATIONAL ACADEMY OF SCIENCES USA*, vol. 102, no. 0027-8424, pp. 13544-13549, 2005.

[9] R.K. Curtis,M. Oresic, and A. Vidal-Puig, "Pathways to the analysis of microarray data," *TRENDS in Biotechnology*, vol. 23, pp. 429-435, 2005.

[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., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *PROCEEDINGS- NATIONAL ACADEMY OF SCIENCES USA*, vol. 102, pp. 15545-15550, 2005.

[11] C. Cheadle, M. P. Vawter, W. J.Freed, K. G. Becker., "Analysis of Microarray Data Using Z Score

Transformation," *JOURNAL OF MOLECULAR DIAGNOSTICS*, vol. 5, pp. 73-81, 2003.

[12] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, B. R. Conklin., "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways," *NATURE GENETICS*, vol. 31, pp. 19-93, 2002.

[13] GeneSifter® Analysis Edition. [Online]. http://www.genesifter.net

[14] R. Pandey, R. K. Guru, and D. W. Mount, "Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data," *BIOINFORMATICS*, vol. 20, pp. 2156-2158, 2004.

[15] T. Zhang, X. Li, Q. Wang, J. Xu, H. Yu, J. Zhu, H. Wang, C. Wang, E. Topol, Q. Wang, and S.Rao Z. Guo, "Towards precise classification of cancers based on robust gene functional expression profiles," *BMC Bioinformatics*, vol. 6, p. 58, 2005.

[16] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J. Vert., "Classification of microarray data using gene networks," *BMC Bioinformatics*, vol. 8, p. 35, 2007.

[17] X. Chen, and L. Wang., "Integrating Biological Knowledge with Gene Expression Profiles for Survival Prediction of Cancer," *Journal of Computational Biology*, vol. 16, pp. 265–278, 2009.

[18] J. Su, B.J. Yoon, and E.R. Dougherty., "Accurate and Reliable Cancer Classification Based on Probabilistic Inference of Pathway Activity," *PLoS One*, vol. 4, pp. 503-511, 2009.

[19] F. Tai and W. Pan., "Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms," *BIOINFORMATICS*, vol. 23, pp. 1775-1782, 2007.

[20] H.Y. Chuang, E. Lee, Y.T. Liu, D. Lee, and T. Ideker., "Network-based classification of breast cancer metastasis," *Molecular systems biology*, vol. 3, p. 140, 2007.

[21] T. Hwang, and T. Park., "Identification of differentially expressed subnetworks based on multivariate ANOVA," *BMC bioinformatics*, vol. 10, p. 128, 2009.

[22] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft., "When is “nearest neighbor” meaningful?," in *Database Theory-7th International Conference Jerusalem*, 1999, pp. 217-235.

[23] W. S. Noble., "What is a support vector machine?," *NATURE BIOTECHNOLOGY*, pp. 1565-1567, 2006.

[24] D. L. Stirewalt, S. Meshinchi, K. J. Kopecky, W. Fan, E. L. Pogosova-Agadjanyan, J. H. Engel, M. R. Cronk, K. S. Dorcy, A. R. McQuary, D. Hockenbery., "Identification of genes with abnormal expression changes in acute myeloid leukemia," *GENES CHROMOSOMES AND CANCER*, vol. 47, no. 1045-2257, pp. 8-20, 2008.

[25] Gene Expression Omnibus database. [Online]. http://www.ncbi.nlm.nih.gov/geo/

[26] Y. Yao, L. Richman, C. Morehouse, M. de Los Reyes, B.W. Higgs, A. Boutrin, B. White, A. Coyle, J. Krueger, P.A. Kiener, and others, "Type I interferon: potential therapeutic target for psoriasis," *PLoS one*, vol. 3, p. e2737, 2008.

[27] W. A. Freije, F. E. Castro-Vargas, Z. Fang, S. Horvath, T. Cloughesy, L. M. Liau, P. S. Mischel, S. F. Nelson., "Gene Expression Profiling of Gliomas Strongly Predicts Survival," *CANCER RESEARCH*, vol. 64, no. 0008-5472, pp. 6503-6510, 2004.