Improving classification accuracy of response in leukaemia treatment using feature selection over pathway segmentation

Zena M. Hira [1,*], Duncan F. Gillies [1] and Edward Curry [2]

[1] Department of Computer Science, Imperial College London, South Kensington Campus, London SW7 2AZ.

[2] Department of Surgery & Cancer, The Commonwealth Building, The Hammersmith Hospital, London W12 0NN

ABSTRACT

Motivation: Many people die every year from leukaemia. Some of them respond to treatment, and some of them not. This study investigates whether there is any relationship between response to treatment and features drawn from the measured methylation profiles of a set of patients. Such features could potentially be used to predict the outcome of a putative treatment regime.

Results: Using AdaBoost with decision trees as weak classifiers, we managed to identify two pathways that affect classification of response and progression in blood cancer with 0.988 accuracy. We also identified a gene whose presence or absence from the dataset can drop classification accuracy from 0.988 to random.

Conclusion: We identified one gene that with 99% accuracy can predict response to treatment. We were also able to identify a list of genes from the same dataset that can predict response with 0.94% accuracy.

Contact: zena.hira09@imperial.ac.uk

1 INTRODUCTION

The latest worldwide cancer statistics, provided by GLOBOCAN 2012, have shown that approximately 14.1 million people suffer from cancer in 2012. The number is expected to rise to 24 million in 20 years time. Leukaemia is the 11th most common cancer for both sexes and it accounts for 2.5% of all cancers with $\approx$ 352,000 new cases worldwide every year[1]. The exact cause of leukaemia is not yet known but it is believed that is related to both environmental and inherited factors. This type of cancer originates in the bone marrow and results in a high number of leukaemia cells (abnormal white blood cells which are not fully developed) (National Cancer Institute, 2013). Approximately 63% $\sim$ 64% patients survive for one year. This rate however drops to 44% after five years and 32% $\sim$ 33% after ten years. There are four major types of leukaemia:

1. Acute Lymphoblastic Leukaemia (ALL): Starts in abnormal lymphoid stem cells and progresses very quickly. (Weinblatt, 2014)
2. Acute Myelogenous Leukaemia (AML):Starts in abnormal myeloid stem cells and develops quickly. (Seiter, 2014)
3. Chronic Lymphocytic Leukaemia (CLL): Starts in abnormal lymphoid stem cells and take months or even years to develop. (Shen *et al.*, 2007)
4. Chronic Myelogenous Leukaemia (CML): CML starts in abnormal myeloid stem cells and develops slowly. (Besa, 2014)

AML, CLL and CML are mainly adult cancers. They are very rarely encountered in children. ALL however is a cancer that is very common in children. Acute leukaemia usually develops quickly and worsens in some weeks unless treated, in contrast to chronic forms of leukaemia that progress very slowly and can be left untreated for months or years.

In this paper we are concerned with the methylation profiles of 91 samples of Chronic Myelogenous Leukaemia (CML), some of which respond to treatment and some not. Even though a lot of progress has been made over the years for the treatment of CML leukaemia, 30% to 35% of patients do not responding to treatment (Carella *et al.*, 2013). A drug called Imatinib, a tyrosine-kinase inhibitor, is the first line of treatment for CML and was introduced in 1988 (Carroll *et al.*, 1997; Smith and Shah, 2011).

Due to the high dimensionality of the dataset we had to find a way of selecting which features to analyse otherwise the analysis would not be possible. We used pathway information from the ConsensusPath database (Kamburov *et al.*, 2013, 2011, 2009; Pentchev *et al.*, 2010) in order to perform a feature selection method for which we initially split the dataset in pathways and then analyse each pathway individually in order to see which pathway gives us more accurate classification. The ConsensusPath database integrates different types of information including: protein interactions, genetic interactions signalling, metabolism, gene regulation and drug target interactions in humans. These are taken from a number of databases including Reactome, KEGG,

*to whom correspondence should be addressed

[1] http://www.cancerresearchuk.org/

HumanCyc, PID, BioCarta. This way we were able to separate the dataset in different groups using prior knowledge.

We used two different methods for linear and non-linear dimensionality reduction methods: PCA and Manifold - Isomap, in order to remove redundant features from the pathway datasets. Then we used AdaBoost with decision trees as weak classifiers as a way of classifying our results since boosting techniques can help with reducing the bias in supervised learning by being less susceptible to the overfitting problem than other learning algorithms (Kearns, 1998).

2 RELATED WORK

Recent improvement in molecular biology technology allowed the analysis of DNA methylation sites and profiling of cells in the whole genome (Schumacher *et al.*, 2006). Methylation is believed to be closely related to gene expression (Aran *et al.*, 2013) and DNA methylation sites have been increasingly found to be involved in processes such as cancer (Laura, 2008; Levenson and Melnikov, 2012). Methylation biomarkers have often been associated with treatment to cancer and response as shown in some clinical studies (Maier *et al.*, 2005; Baylin, 2005). Machine learning has been widely used on biological data with increasing success (Wang *et al.*, 2005; Osareh and Shadgar, 2010; Liu *et al.*, 2009). Methylation data have only recently started being analysed using machine learning (Ruan *et al.*, 2012a,b; Wilhelm, 2014) so there is still a lot to be discovered.

Our aim was to combine methylation data with prior knowledge in our new feature selection approach. Prior knowledge has been used before in classification of microarray expression using information about the genes (Brown *et al.*, 2000; Guan *et al.*, 2009), pathway information (Hira *et al.*, 2014) or GO terms (Chen and Xu, 2004; Kustra and Zagdanski, 2010; Cheng *et al.*, 2004; Chen and Wang, 2009). Little has been done in terms of predicting response to cancer treatment using methylation data and prior knowledge.

3 METHODS

3.1 Dataset

DNA methylation is an epigenetic mark that potentially has a regulatory role in gene expression (Bock, 2012). It can be used to identify biomarkers for a number of diseases, including cancer, since it can provide information about environmental exposures (Walker and Ho, 2012). Methylation occurs at CpG islands and it means that a methyl group is added to a cytosine residue to convert it to 5-methylcytosine. A CpG site is a place on the linear sequence of the bases of the DNA that has a cytosine and guanine separated by only one phosphate. Methylation of these sites that are in the promoters of genes can affect their expression and lead to their silencing, a feature found in a number of human cancers (Jones and Laird, 1999). In our methylation dataset there were 429 231 probes with 91 samples, 60 of them were responsive to blood cancer treatment while 31 were not.

3.2 Approach

Given its large number of probes, the dataset is very difficult to analyse as it is. In order to proceed with the analysis, we decided to split the dataset based on the pathways found in the ConsensusPath database (Kamburov *et al.*, 2013, 2011, 2009; Pentchev *et al.*, 2010). Genes that belong in the same pathway were put in the same gene set. Therefore the dataset was divided in 2072 smaller sets between approximately 100 and 2000 genes each, as shown in figure 1. Our method works as follows:

- Split the dataset into pathways
- Apply dimensionality reduction
- Classify progression and response to treatment

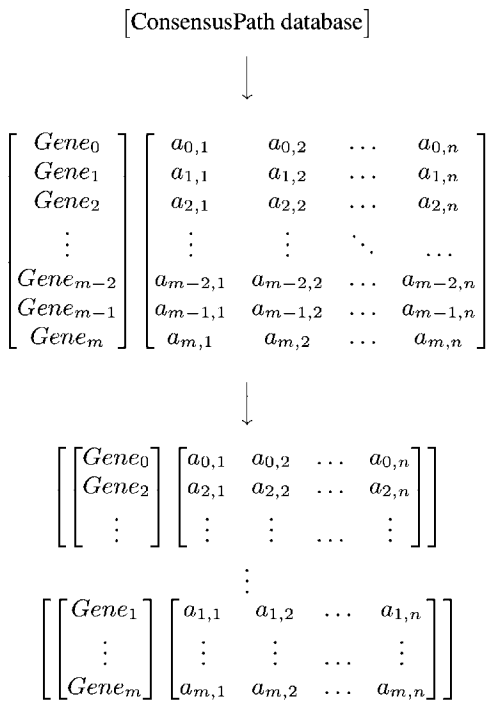

Fig. 1. The original methylation dataset, split into smaller sets based pathway information

The methods used for dimensionality reduction were Principal Components Analysis (PCA) and Manifold projection (ISOMAP).

3.3 Theoretical Background

3.3.1 Principal Components Analysis (PCA) Principal Components Analysis (Bishop, 2006) can be used to reduce the dimensionality of the data. It finds the principal variations among the data and ignores the smaller ones. Simply put, it compares data in terms of similarities and differences. The simplification of data should be done in a way where the important features are not lost. There are a number of steps to be done in order to apply PCA.

1. Subtract the mean for each data dimension of the data (to get the *mean-adjusted data*). Mean is given by:
$$\bar{x}[i] = 1/N \sum_{n=1}^{N} X[i,n] \quad (1)$$
2. Calculate the Covariance matrix.
$$cov(X,Y) = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})/(N-1) \quad (2)$$
3. Calculate the eigenvectors and eigenvalues of the covariance matrix .
4. Order the eigenvectors from highest to lowest and chose the ones with the highest values i.e. the most significant ones. The number of vectors

selected will represent the number of dimensions the new dataset will have.

5. Construct a *feature vector* which is a matrix of vectors

$$featureVector = (eigenVector_1, eigenVector_2, ..., eigenVector_n) \quad (3)$$

6. Derive the new dataset by transposing and multiplying the feature vector and the mean adjusted data.

$$finalDataset = featureVector^T * MeanAdjustedData^T \quad (4)$$

This will return the dataset in the axis system defined by the eigenvectors

3.3.2 Manifold - Isomap The manifold learning algorithm is used for non-linear dimensionality reduction (Cayton, 2005). Manifold learning generally works by taking inputs from a higher dimensional space and embeds them to a lower one while preserving their characteristics. It assumes that all data points are lying close to or on a manifold and it can be thought as a generalised principal components analysis (PCA) that can capture non-linear relations. Isomap, (Tenenbaum *et al.*, 2000) short for Isometric Mapping, was one of the first approaches to manifold construction and is an extension to *Kernel PCA*. The Isomap algorithm works as follows:

1. Determine the neighbours of each data point: For all points in a fixed radius, find the k nearest points (k - Isomap) or the closest points based on distance (ϵ-Isomap)
2. Construct the neighbourhood graph: Points are connected to each of their k nearest points with the edge length set to their Euclidean distance.
3. Find the shortest path between all the nodes on the graph using a graph algorithm (*Dijkstra* or *Floyd-Warshall*) to construct the matrix of pairwise geodesic distances between different points.
4. Construct the lower dimensionality mapping. This is the same procedure as classical MDS. Generally another matrix Θ is constructed using:

$$\Theta = -\frac{1}{2} H \Delta^2 H \quad (5)$$

where Δ is the matrix of geodesic distances;
and H is the centering matrix:

$$H = I_n - \frac{1}{N} U_N \quad (6)$$

where U_N is an $N \times N$ matrix of 1's;
and I_n is the identity matrix of size n

5. Calculate the eigenvalues of Θ: Let λ_k be the k^{th} eigenvalue and v_k be the k^{th} eigenvector. The k^{th} component of the embedding Π is constructed by setting it to $\sqrt{\lambda_k} v_k$.

$$\Pi = \begin{pmatrix} \sqrt{\lambda_1} v_1 \\ \sqrt{\lambda_2} v_2 \\ \sqrt{\lambda_3} v_3 \\ \vdots \\ \sqrt{\lambda_d} v_d \end{pmatrix}$$

3.3.3 AdaBoost with Decision Trees as the weak classifiers We used AdaBoost for the classification of progression or non - progression. Adaboost is an ensemble method for classifying that can improve the quality of the output. Instead of using only one machine learning method, ensemble methods use a combination of weak classifiers. Each classifier is an individual hypothesis about the data. A set of classifiers is constructed and new data is classified by taking a vote on the classifiers' predictions. For ensemble methods to work as accurately as possible they need to be diverse. Diversity means that they make different errors in the newly classified data (Dietterich, 2000). The AdaBoost classifier structure is shown in figure 2. Boosting works by having a number of weak classifiers i.e. classifiers whose performance is marginally better than random and combining them in a way that creates a strong classifier. AdaBoost starts with one classifier fitted to the dataset and then it creates more copies of it, that are again applied on the data. The classifier's weights are adjusted according to the accuracy of the result and normally on subsequent runs of the program they are modified so that they can accommodate the most difficult cases (Freund and Schapire, 1997). We use the AdaBoost-SAMME algorithm (Hastie *et al.*, 2009), a multiclass version of the original algorithm. Decision trees were used as the weak classifiers. A decision trees takes input tuples of the form $(X, Y) = (x_1, x_2, ..., x_k, Y)$ and it creates rules based on $(x_1, x_2, ..., x_k)$ so that the target Y can be classified correctly. The tree is constructed by splitting the inputs recursively (recursive partitioning) and it ends when the subset at a node has items with the same label or when the accuracy can no longer be improved using the Gini Impurity shown in equation 7. The Gini Impurity measures how often a random element can be labelled incorrectly if a random label was assigned to it based on the distribution of labels on the set.

$$I_G(f) = \sum_{i=1}^{m} f_i(1-f_i) = \sum_{i=1}^{m} (f_i - {f_i}^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} {f_i}^2 = 1 - \sum_{i=1}^{m} {f_i}^2 \quad (7)$$

The decision tree algorithm we used is the Classification And Regression Tree (CART) (Breiman *et al.*, 1984) which works with both categorical and numerical target variables. It creates the tree with the features that give the biggest information gain at each node. In figure 2 $h_1(x)$, $h_i(x)$, $h_k(x)$ denote the weak classifiers. AdaBoost, PCA and Isomap were implemented in the sklearn (Pedregosa *et al.*, 2011) package for Python.

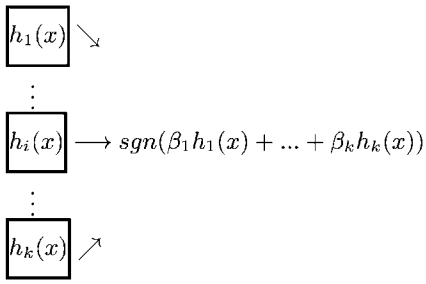

Fig. 2. AdaBoost Classifier with decision trees as the weak classifiers

4 RESULTS

4.1 Pathways

We performed classification on all pathway sets in order to see how well each of them could classify blood cancer progression. We established that in most cases using dimensionality reduction worsens the results. This is because the datasets are already small enough and performing dimensionality reduction loses most of the information the sets have. We isolated 2 pathways that were giving us accuracy of 0.9888. To get the accuracy we used 10 fold stratified cross validation.

We focused more on the most accurate pathways since their accuracy is significantly higher. We plotted the Receiver operating

Table 1. Pathways with the highest scores

Pathway Name	Accuracy	
Regulation of KIT signaling	0.9888	0.0011
Signaling events mediated by Stem cell factor receptor (c-Kit)	0.9888	0.0011
Superpathway of D-myo-inositol (1,4,5)-trisphosphate metabolism	0.8244	0.0176
D-myo-inositol (1,4,5)-trisphosphate metabolism	0.813	0.0098
3-phosphoinositide degradation	0.79	0.0136

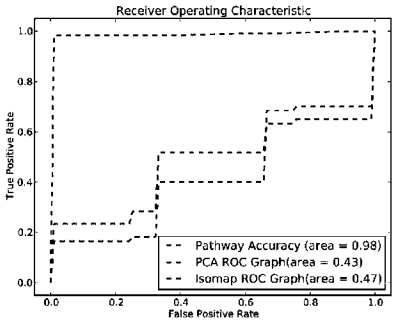

Fig. 3. Regulation of KIT signalling pathway set, comparison of the pathway set, with and without dimensionality reduction (PCA and Isomap).

characteristic (ROC) curves for those two pathways. We compared the accuracy of the Regulation of KIT signalling pathway set using two different methods of dimensionality reduction and without any dimensionality reduction. The linear algorithm for finding the significant features is PCA and the non-linear Manifold - Isomap. The graph is shown in figure 3. For Signalling events mediated by Stem cell factor receptor (c-Kit) pathway set the ROC curve is shown in 4

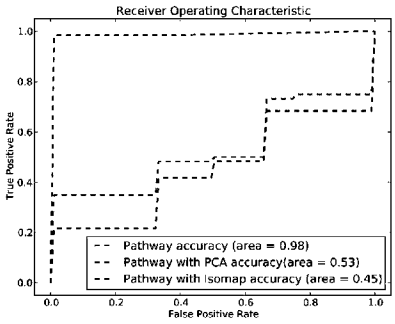

Fig. 4. Signalling events mediated by Stem cell factor receptor (c-Kit) pathway set, comparison of the pathway set, with and without dimensionality reduction (PCA and Isomap).

We show the ROC curves for two other pathways that do not perform so well (figures 5, 6 and 7).

We used the two higher-scoring pathways to create sets that only have a percentage of the genes in, to check if this will affect the

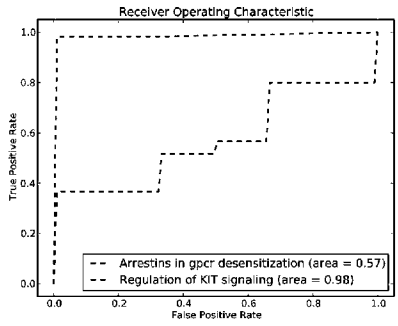

Fig. 5. Comparison between Regulation of KIT signalling and Arrestins in gpcr desensitization pathway sets

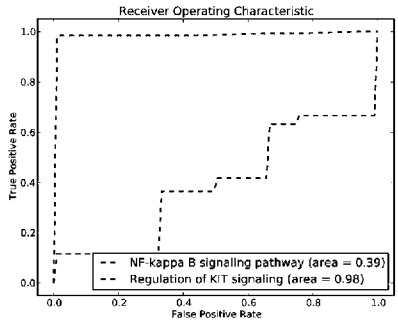

Fig. 6. Comparison between Regulation of KIT signalling and NF-kappa B signalling pathway - Homo sapiens

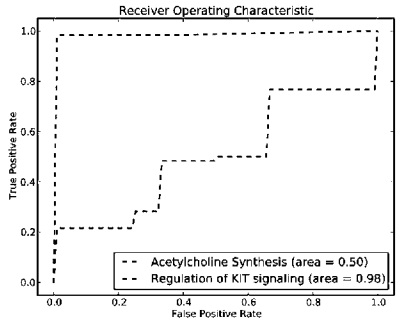

Fig. 7. Comparison between Regulation of KIT signalling and Acetylcholine Synthesis

accuracy. We concluded that most of the time removing genes from the pathway sets does affect the classification accuracy. There are occasions were the accuracy was not affected. That is because accuracy depends on which genes we removed. This is shown in supplementary information. In addition random gene sets were created to verify the effectiveness of pathways. We constructed the graphs for the highest-scored pathways against random gene sets. It is shown that random gene sets have far less accuracy. These results are included in the supplementary information.

4.2 Gene SH2B3

Investigating further, we studied how the AdaBoost was constructing the classifier models for the data. We found out that the CpG Island cg00056489, which translates to gene *SH2B3* or *SH2B adaptor protein 3*, was the gene that most of the modelling was based on. In fact removing this gene only out of the pathway set reduced the classification accuracy to random (≈ 0.5) from the initial 0.99 that was before.

4.3 Gene Set

We searched for other genes in the whole dataset that had an important impact on the classification. We managed to identify another set of genes that can classify response to treatment with accuracy of 0.94. To obtain this set we went through the pathways applying AdaBoost. We picked the pathways that could get an accuracy of more than a threshold. Having a closer look at how AdaBoost builds the decision classifiers we could see which features (genes) were the most important when building the tree. We filtered out the features that were more important in constructing the classifier. We constructed classifiers for each of those combinations of values and we picked the ones where the most features were taken into consideration when building the classifier. The algorithm is shown in figure 8. $AccuracyThreshold$ had values between $0.7 - 0.9$ and $ImportanceThreshold$ had different values ranging from 0.003 to 0.5 We found the list of genes shown in table 2.

```
Data: Methylation Data
for p ∈ PathwaySet do
    if accuracy(p) ≥ AccuracyThreshold then
        for feature ∈ DecisionTree(p) do
            if importance(feature) ≥ ImportanceThreshold then
                GeneSet ← feature;
            end
        end
    end
end
```

Fig. 8. Gene Selection Algorithm

When removing gene *SH2B adaptor protein 3* which is the one that we identified before, the accuracy drops to 0.94. *SH2B adaptor protein 3* seems to be a direct way of classifying the response to treatment. This gene set seems to be equally important in terms of accuracy. Removing any other gene from this dataset while SH2B3 is not present reduces the accuracy as shown in table 3.

Table 2. Gene Set

Gene Name	Cg Island	Functional Annotation
INPP5A	cg11930406	inositol polyphosphate-5-phosphatase, 40kDa
INPP5A	cg10762214	inositol polyphosphate-5-phosphatase, 40kDa
INPP5A	cg07876930	inositol polyphosphate-5-phosphatase, 40kDa
INPP5A	cg23714705	inositol polyphosphate-5-phosphatase, 40kDa
INPP5A	cg17859552	inositol polyphosphate-5-phosphatase, 40kDa
INPP5A	cg12507869	inositol polyphosphate-5-phosphatase, 40kDa
INPP5A	cg03149567	inositol polyphosphate-5-phosphatase, 40kDa
INPP5A	cg24608475	inositol polyphosphate-5-phosphatase, 40kDa
INPP5A	cg23009449	inositol polyphosphate-5-phosphatase, 40kDa
INPP5A	cg10509185	inositol polyphosphate-5-phosphatase, 40kDa
INPP5B	cg03523189	inositol polyphosphate-5-phosphatase, 75kDa
INPP5F	cg11613559	inositol polyphosphate-5-phosphatase F
INPP5F	cg20365618	inositol polyphosphate-5-phosphatase F
INPP5F	cg07679322	inositol polyphosphate-5-phosphatase F
INPP5F	cg02857557	inositol polyphosphate-5-phosphatase F
INPP5F	cg02722214	inositol polyphosphate-5-phosphatase F
IMPAD1	cg03732295	inositol monophosphatase domain containing 1
INPP1	cg17516156	inositol polyphosphate-5-phosphatase J
INPP5J	cg27324619	inositol polyphosphate-5-phosphatase J
ITPKB	cg04242667	inositol 1,4,5-trisphosphate 3-kinase B
ITPKB	cg14711690	inositol 1,4,5-trisphosphate 3-kinase B
SH2B3	cg00056489	SH2B adaptor protein 3
SYNJ2	cg02118886	synaptojanin 2
SYNJ2	cg22242614	synaptojanin 2
–	cg01195672	–
–	cg07655693	–
–	cg16687447	–
–	cg16832975	–

Table 3. Removing Genes from Set

Gene Name	Cg Island	Accuracy
INPP5A	cg11930406	0.9678
INPP5A	cg10762214	0.9333
INPP5A	cg07876930	0.9456
INPP5A	cg23714705	0.9556
INPP5A	cg17859552	0.9344
INPP5A	cg12507869	0.9344
INPP5A	cg03149567	0.9556
INPP5A	cg24608475	0.9667
INPP5A	cg23009449	0.9456
INPP5A	cg10509185	0.9456
INPP5B	cg03523189	0.9678
INPP5F	cg11613559	0.9344
INPP5F	cg20365618	0.9444
INPP5F	cg07679322	0.9678
INPP5F	cg02857557	0.9678
INPP5F	cg02722214	0.9678
IMPAD1	cg03732295	0.9567
INPP1	cg17516156	0.9456
INPP5J	cg27324619	0.9122
ITPKB	cg04242667	0.9567
ITPKB	cg14711690	0.9567
SYNJ2	cg02118886	0.9244
SYNJ2	cg22242614	0.9567
–	cg01195672	0.9456
–	cg07655693	0.9556
–	cg16687447	0.9244
–	cg16832975	0.9556

5 DISCUSSION

We identified gene SH2B3 to be controlling the response to CML leukaemia treatment. *SH2B adapter protein 3* is a protein that in humans is encoded by the SH2B3 gene (Motto *et al.*, 1996; Hendricks-Taylor *et al.*, 1997). Its role is to be involved in a range of signalling activities by growth factor and cytokine receptors. It is a member of the family of tyrosine kinase adapter proteins (Ahmed *et al.*, 1999), the high-affinity cell surface receptors for many polypeptide growth factors, cytokines, and hormones (Robinson *et al.*, 2000), which are shown to be involved with the progression

of many types of cancer. The possibility of manipulating receptor tyrosine kinase signalling in order to prevent cancer or enhance cancer therapy was explored previously in (Zwick *et al.*, 2001). It is a key protein for the negative regulator of cytokine signalling and plays a critical role in hematopoiesis. This kind of cells are very much related with leukaemia (Sachs, 1996; National Cancer Institute, 2013). More over SH2B3 has already been identified as a predisposition gene to Acute Lymphoblastic Leukaemia (ALL) (Willman, 2013).

From the set of genes that can also predict response very accurately (0.94), inositol polyphosphate-5-phosphatase has already been associated with leukaemia in (Mengubas *et al.*, 1994). In addition it is associated with *SH2* since it encodes a protein in that domain. The protein is related to hematopoietic cells and its movement from the cytosol to the plasma membrane is mediated by tyrosine phosphorylation (Liu *et al.*, 1998).

Synaptojanin was also found in the set which belongs to the inositol-polyphosphate 5-phosphatase family that has previously been associated with hairy cell leukaemia, a chronic mature B-cell leukemia characterized by malignant B cells that have typical hairy protrusions (Spaenij-Dekking *et al.*, 2003).

6 CONCLUSION

We present a way of analysing big datasets by segmenting them based on prior pathway information. Analysing the pathways separately can give us an idea as to how a disease is related to a pathway and which biological mechanisms are involved. We used AdaBoost for classification since it can remove bias in supervised learning. It also reduces computational time and complexity. Indeed analysing the whole dataset it would not be possible.

We identified gene SH2B3 that belongs in the family of tyrosine kinase adapter proteins that has previously been associated with leukaemia. SH2B3 was also shown to be related to predisposition to ALL. We also identified a set of genes that can be almost as accurate as SH2B3. Most of the genes in that list belong to the family of inositol polyphosphate-5-phosphatase which are associated with *SH2*. Further experimentation and analysis must be performed to determine whether these results can be used to define an effective biomarker in a clinical setting in the battle against leukaemia.

ACKNOWLEDGEMENT

Funding: Imperial College Student Scholarship

REFERENCES

Ahmed, Z., Smith, B. J., Kotani, K., Wilden, P., and Pillay, T. S. (1999). Aps, an adapter protein with a ph and sh2 domain, is a substrate for the insulin receptor kinase. *Biochem J*, **341 (Pt 3)**, 665–8.

Aran, D., Sabato, S., and Hellman, A. (2013). DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biology*, **14**(3), R21+.

Baylin, S. B. (2005). Dna methylation and gene silencing in cancer. *Nature clinical practice. Oncology*, **2 Suppl 1**, S411.

Besa, E. C. (2014). Chronic myelogenous leukemia.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Bock, C. (2012). Analysing and interpreting dna methylation data. *Nat Rev Genet*, **13**(10), 705–19.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., and Haussler, D. (2000). Knowledge–based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, **97**(1), 262–267.

Carella, A. M., Branford, S., Deininger, M., Mahon, F. X., Saglio, G., Eiring, A., Khorashad, J., O'Hare, T., and Goldman, J. M. (2013). What challenges remain in chronic myeloid leukemia research? *Haematologica*, **98**(8), 1168–1172.

Carroll, M., Ohno-Jones, S., Tamura, S., Buchdunger, E., Zimmermann, J., Lydon, N. B., Gilliland, D. G., and Druker, B. J. (1997). Cgp 57148, a tyrosine kinase inhibitor, inhibits the growth of cells expressing bcr-abl, tel-abl, and tel-pdgfr fusion proteins. *Blood*, **90**(12), 4947–4952.

Cayton, L. (2005). Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep.*

Chen, X. and Wang, L. (2009). Integrating biological knowledge with gene expression profiles for survival prediction of cancer. *Journal of Computational Biology*, **16**(2), 265–278.

Chen, Y. and Xu, D. (2004). Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae. *Nucleic Acids Res*, **32**(21), 6414–6424.

Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., Kulp, D., and Siani-Rose, M. (2004). A knowledge-based clustering algorithm driven by gene ontology. *J Biopharm Stat*, **14**(3), 687–700.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, pages 1–15, London, UK, UK. Springer-Verlag.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**(1), 119–139.

Guan, P., Huang, D., He, M., and Zhou, B. (2009). Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method. *Journal of Experimental & Clinical Cancer Research*, **28**(1), 103+.

Hastie, T., Rosset, S., Zhu, J., and Zou, H. (2009). Multi-class AdaBoost. *Statistics and Its Interface*, **2**(3), 349–360.

Hendricks-Taylor, L. R., Motto, D. G., Zhang, J., Siraganian, R. P., and Koretzky, G. A. (1997). Slp-76 is a substrate of the high affinity ige receptor-stimulated protein tyrosine kinases in rat basophilic leukemia cells. *J Biol Chem*, **272**(2), 1363–7.

Hira, Z. M., Trigeorgis, G., and Gillies, D. F. (2014). An algorithm for finding biologically significant features in microarray data based on ¡italic¿a priori¡/italic¿ manifold learning. *PLoS ONE*, **9**(3), e90562.

Jones, P. A. and Laird, P. W. (1999). Cancer epigenetics comes of age. *Nature genetics*, **21**(2), 163–167.

Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2009). Consensuspathdba database for integrating human functional interaction networks. *Nucleic Acids Research*, **37**(suppl 1), D623–D628.

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. (2011). Consensuspathdb: toward a more complete picture of cell biology. *Nucleic Acids Research*, **39**(suppl 1), D712–D717.

Kamburov, A., Stelzl, U., Lehrach, H., and Herwig, R. (2013). The consensuspathdb interaction database: 2013 update. *Nucleic Acids Research*, **41**(D1), D793–D800.

Kearns, M. (1998). Thoughts on hypothesis boosting.

Kustra, R. and Zagdanski, A. (2010). Data-fusion in clustering microarray data: Balancing discovery and interpretability. *IEEE/ACM Trans. Comput. Biology Bioinform.*, **7**(1), 50–63.

Laura, B. (2008). Epigenomics: The new tool in studying complex diseases. *Nature Education*, **1**(178).

Levenson, V. V. and Melnikov, A. A. (2012). Dna methylation as clinically useful biomarkerslight at the end of the tunnel. *Pharmaceuticals*, **5**(1), 94–113.

Liu, Q., Shalaby, F., Jones, J., Bouchard, D., and Dumont, D. J. (1998). The sh2-containing inositol polyphosphate 5-phosphatase, ship, is expressed during hematopoiesis and spermatogenesis. *Blood*, **91**(8), 2753–9.

Liu, Q., Sung, A. H., Chen, Z., Liu, J., Huang, X., and Deng, Y. (2009). Feature selection and classification of maqc-ii breast cancer and multiple myeloma microarray gene expression data. *PLoS ONE*, **4**(12), e8250.

Maier, S., Dahlstroem, C., Haefliger, C., Plum, A., and Piepenbrock, C. (2005). Identifying DNA methylation biomarkers of cancer drug response. *Am J Pharmacogenomics*, **5**(4), 223–32+.

Mengubas, K., Jabbar, S., Nye, K., Wilkes, S., Hoffbrand, A., and Wickremasinghe, R. (1994). Inactivation of calcium ion-regulating inositol polyphosphate second messengers is impaired in subpopulations of human leukemia cells. *Leukemia*,

8(10), 1718–1725.

Motto, D. G., Musci, M. A., Ross, S. E., and Koretzky, G. A. (1996). Tyrosine phosphorylation of grb2-associated proteins correlates with phospholipase c gamma 1 activation in t cells. *Mol Cell Biol*, **16**(6), 2823–9.

National Cancer Institute (2013). What you need to know about leukemia.

Osareh, A. and Shadgar, B. (2010). Machine learning techniques to diagnose breast cancer. In *Health Informatics and Bioinformatics (HIBIT), 2010 5th International Symposium on*, pages 114–120.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Pentchev, K., Ono, K., Herwig, R., Ideker, T., and Kamburov, A. (2010). Evidence mining and novelty assessment of proteinprotein interactions with the consensuspathdb plugin for cytoscape. *Bioinformatics*, **26**(21), 2796–2797.

Robinson, D. R., Wu, Y.-M., and Lin, S.-F. (2000). The protein tyrosine kinase family of the human genome. *Oncogene*, **19**(49), 5548–5557.

Ruan, J., Jahid, M. J., Gu, F., Lei, C., Huang, Y.-W., Hsu, Y.-T., Mutch, D. G., Chen, C.-L., Kirma, N. B., and Huang, T. H. (2012a). Network-based classification of recurrent endometrial cancers using high-throughput dna methylation data. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, BCB '12, pages 418–425, New York, NY, USA. ACM.

Ruan, J., Jahid, M., Gu, F., Lei, C., Huang, Y.-W., Hsu, Y.-T., Mutch, D., Chen, C.-L., Kirma, N., and Huang, T. (2012b). Network-based classification of recurrent endometrial cancers using high-throughput dna methylation data. pages 418–425. cited By (since 1996)0.

Sachs, L. (1996). The control of hematopoiesis and leukemia: from basic biology to the clinic. *Proc Natl Acad Sci U S A*, **93**(10), 4742–9.

Schumacher, A., Kapranov, P., Kaminsky, Z., Flanagan, J., Assadzadeh, A., Yau, P., Virtanen, C., Winegarden, N., Cheng, J., Gingeras, T., and Petronis, A. (2006). Microarray-based DNA methylation profiling: technology and applications. *Nucleic acids research*, **34**(2), 528–542.

Seiter, K. (2014). Acute myelogenous leukemia.

Shen, L., Au, W.-Y., Guo, T., Wong, K.-Y., Wong, M. L., Tsuchiyama, J., Yuen, P.-W., Kwong, Y.-L., Liang, R. H., and Srivastava, G. (2007). Proteasome inhibitor bortezomib-induced apoptosis in natural killer (nk)–cell leukemia and lymphoma: an in vitro and in vivo preclinical evaluation. *Blood*, **110**(1), 469–470.

Smith, C. C. and Shah, N. P. (2011). Tyrosine kinase inhibitor therapy for chronic myeloid leukemia: Approach to patients with treatment-naive or refractory chronic-phase disease. *ASH Education Program Book*, **2011**(1), 121–127.

Spaenij-Dekking, E. H. A., Van Delft, J., Van Der Meijden, E., Hiemstra, H. S., Falkenburg, J. H. F., Koning, F., Drijfhout, J. W., and Kluin-Nelemans, J. C. (2003). Synaptojanin 2 is recognized by hla class ii-restricted hairy cell leukemia-specific t cells. *Leukemia*, **17**(12), 2467–73.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, **290**(5500), 2319–2323.

Walker, C. L. and Ho, S.-m. (2012). Developmental reprogramming of cancer susceptibility. *Nat Rev Cancer*, **12**(7), 479–86.

Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F., and Mewes, H. W. (2005). Gene selection from microarray data for cancer classificationa machine learning approach. *Computational Biology and Chemistry*, **29**(1), 37 – 46.

Weinblatt, M. E. (2014). Pediatric acute myelocytic leukemia.

Wilhelm, T. (2014). Phenotype prediction based on genome-wide dna methylation data. *BMC Bioinformatics*, **15**, 193.

Willman, C. L. (2013). Sh2b3: a new leukemia predisposition gene. *Blood*, **122**(14), 2293–2295.

Zwick, E., Bange, J., and Ullrich, A. (2001). *Receptor Tyrosine Kinase Signalling as a Target for Cancer Intervention Strategies*. Journal of Endocrinology Limited.

1 Supplementary Information

1.1 Partial Pathway Sets

Pathway accuracy depends on which genes were removed from the set as shown in figure 1 and figure 2. Sometimes the accuracy is not affected as shown in figure 3 and figure 4.

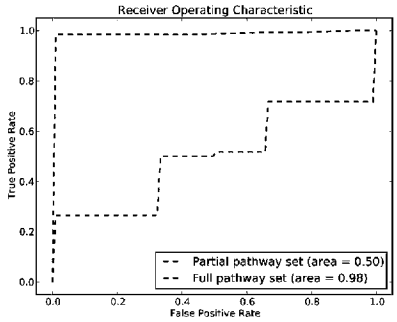

Figure 1: 50% of the pathways is removed

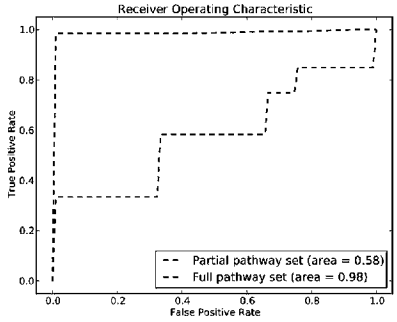

Figure 2: 60% of the pathways is removed

1.2 Random Sets

In addition to prove that pathways are effective in classification we created random sets which have the same size as the pathways and compared the classification results between them and the two pathways that perform best. The results are shown in figure 5, figure 6 and figure 7.

Comparing the accuracy of all the random sets, after 10 different runs we saw that the accuracies vary between 0.4 and 0.7. Their variance is either

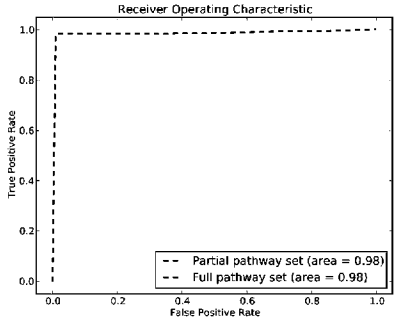

Figure 3: 80% of the pathways is removed

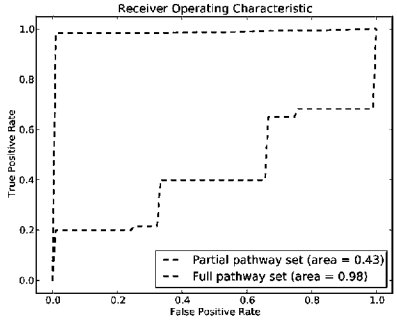

Figure 4: 80% of the pathways is removed

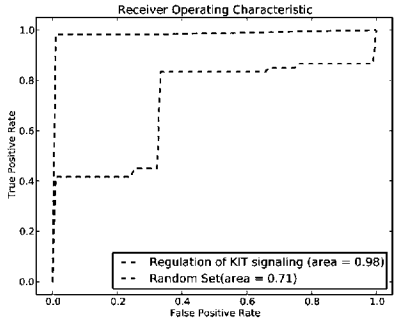

Figure 5: Comparison between Regulation of KIT signalling and a random pathway set of 1170 genes

equal or much greater than the variance of the pathway sets. The variance of the two pathways is 0.001111111 while for the random sets variance is between 0.016474074 and 0.025955556. This proves that the pathways are not random collections of genes and they do play an important role in classifying

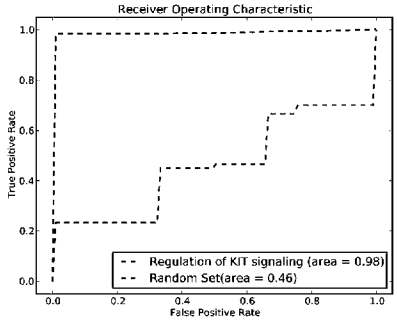

Figure 6: Comparison between Regulation of KIT signalling and a random pathway set of 644 genes

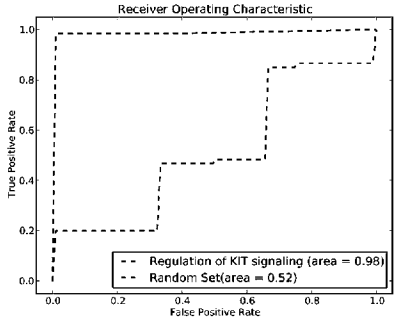

Figure 7: Comparison between Regulation of KIT signalling and a random pathway set of 728 genes

cancer.