Chapter 5 RVMs for Classification of Expression Data

5.1 Introduction

The phenotypic behaviour of a cell is in large part due to the activity of its proteins. These are translated from mRNAs, which are have been transcribed from active genes. There are many levels at which the activity of proteins can be regulated, however it has generally become accepted that measuring mRNA levels gives a good insight into the relative levels of gene activity. The results of simultaneous measurements of large numbers of mRNA levels, made in a single experiment, will be referred to as 'expression data'.

It has become possible to collect expression data systematically using methods such as quantitative PCR (Buck, Harris et al. 1991; Nedelman, Heagerty et al. 1992), array technologies (Schena, Shalon et al. 1995; Lashkari, DeRisi et al. 1997; Shalon 1998) and DNA chips (Guo, Guilfoyle et al. 1994; Hughes, Mao et al. 2001). The availability of complete genomes and fairly complete gene annotation has enabled the construction of DNA probes to represent most expressed mRNAs in a particular population of cells. Many thousands of these probes can be arrayed onto a single microarray or DNA chip, making it possible to capture a snapshot of the expression state of a cell in a single measurement.

Given the availability of this measurement technology, it becomes possible to take snapshots of a range of conditions of a population of cells and, by processing the results, compare the expression levels of different genes under different conditions. Examples include comparing the natural state of the cell with its state during

conditions of metabolic stress, heat shock or disease. These measurements generate large volumes of data that can contain large statistical errors as a result of limitations inherent to the experimental technologies. Errors may also arise from stochastic variations in the different cell populations under study, due to the inherent dynamics of complex systems (for example, see (Guillouzic, L'Heureux et al. 2000; Smolen, Baxter et al. 2001) for discussions of ways in which the dynamics of gene expression are inherently complex systems).

For many cellular processes, we have a fair understanding of the ways groups of genes are co-regulated as a result of biochemical, genetic and other analysis. Expression data gives us the opportunity to systematically extend this understanding to the whole genome, showing previously unknown regulatory relationships. The expectation is that genes which appear to be co-regulated are likely to be involved in the same cellular processes. One way of viewing this data is from the point of view of genes, for example, the level of a gene during sporulation (Chu, DeRisi et al. 1998). Another way is to classify conditions, i.e. to match a particular cellular expression snapshot to a particular cancer condition (Alizadeh, Eisen et al. 2000; Ramaswamy, Tamayo et al. 2001).

The standard method for processing expression data is currently cluster analysis (Eisen, Spellman et al. 1998). This describes the dynamics of expression data as a hierarchical model, either in terms of similar experimental samples given genes, or similar genes given a set of experiments. Many of the major signals that emerge from naïve clustering are strongly correlated with histological features, for example in different areas of the gut (Bates, Erwin et al. 2002), or different developmental stages (Mody, Cao et al. 2001). Sometimes, by selecting sub-trees of genes, sets can be

found that co-segregate with a qualitative or quantitative observation. However, this is far from being an automated task, relying on human concepts of relevance and relatedness.

One practical application of expression data analysis using machine learning techniques has been the classification of cancer cell types from different patients. Different cancer cell types can appear very similar, but may have very different survival rates which may require different treatments (Kihara, Tsunoda et al. 2001; Liefers and Tollenaar 2002; van 't Veer, Dai et al. 2002). In (Ramaswamy, Tamayo et al. 2001) SVMs were used to construct a classifier for the expression patterns of 14 distinct tumour types from 218 samples. Each expression data sample included measurements from 16,063 genes. The SVM-based classifier could then be used to classify the tumour type of any new sample with a high degree of accuracy (78%). They were also able to identify which genes contributed most to the SVM model.

SVMs were able to carry out the above classification as a result of the large expression differences between cancer cell types. A much more difficult problem is to automatically extract information about changes in gene expression as a result of a much subtler difference, such as in response to drug treatment. The subtler effects tend to be masked by the differences between the cell lineages that have been treated.

Perou (Perou, Sorlie et al. 2000) describes a series of experiments measuring 8,102 gene expression levels in human breast cancer cell lines and biopsy samples. One sub-set of the samples are biopsies taken from patients with tumours before and after treatment with the anti-cancer drug, doxorubicin. The expression data for 20 before-after pairs were clustered using hierarchal clustering. All but three of the sample pairs clustered together as siblings in the tree. This indicates that the changes in gene

expression due to treatment are not reliably detected by clustering against the background of the cell lineage. In a further paper (Brenton, Aparicio et al. 2001), cluster-analysis could be used to identify sub-types of breast tumours. However, this required much larger amounts of data and some human intervention. It also was unable to address the issue of what effect the doxorubicin had upon gene expression.

Typically, when modelling expression data, the aim is both to perform some classification task, and also to look at the resulting model and identify genes that contribute to the model, with the hope that they will provide biological insight. In this chapter the use of SVMs and RVMs is explored to extract information about the effects of doxorubicin. We both evaluate whether these methods are able to generate models that can classify micro-arrays into pre- and post-treatment with doxorubicin, and also to decide if the models they produce are consistent with the known biological processes, and therefore could be used in other situations to identify novel relevant genes.

5.2 Cellular Responses to Doxorubicin

Doxorubicin causes cellular apoptosis by several routes. The primary action of doxorubicin activity is due to intercalation into double-stranded DNA. This both prevents the normal UV-repair pathway (nucleotide excision repair) and causes single-strand breaks, both of which lead to an increase in the rate of DNA repair enzyme activation. Normally, topoisomerase II relaxes tension due to supercoiling by scanning DNA. It then breaks one of the two phosphate backbones, allows the DNA to relax and then repairs the resulting single strand break. If instead it binds to an intercalated doxorubicin molecule, the single-strand break is made, but is not repaired. In this case, the topoisomerase II protein remains covalently attached to the broken

strand (Tewey, Rowe et al. 1984). The activity level of the DNA repair response is measured by the cell, and if it increases above a critical threshold, the cell enters an apoptotic response.

Cell lines that are resistant to doxorubicin often share a common set of mutations. Topoisomerase II activity can be severely impaired (Potmesil, Hsiang et al. 1988). This is consistent with the role of this gene in the drug's mechanism of action. Resistant cell lines frequently express multi-drug resistance proteins, such as P-glycoprotein (P-gp), and multi-drug resistance associated protein (MRP) (Grandjean, Bremaud et al. 2001) which expel the drug from the cells. Impaired systems for maintaining levels of small ions, such as Na^+ and K^+ (Lawrence 1988; Lawrence and Davis 1990) also seem to confer a measure of resistance. It is possible that these small ions are required to enhance the stability of the complex between topoisomerase II and the DNA. Resistant cells often have impaired Jun-Fos pathways (Pourquier, Montaudon et al. 1998). During apoptosis, the Jun-Fos transcription factor heterodimer is activated via a signal-cascade. This in turn leads to the altered expression of gene products, activating the signal-cascades mediating the cellular apoptotic pathway. If either Jun or Fos gene is mutated to loss-of-function mutants, then is apoptosis pathway can not activate.

5.3 Generalized Linear Models

Although SVMs have been used successfully to distinguish between tumour types where there have been large numbers of samples available as described above (Ramaswamy, Tamayo et al. 2001), previous implementations guarantee that the SVM will find a solution even if there is insufficient evidence to support it. The training algorithms for SVMs search for the globally 'best' separating hyper-plane,

and give no indication of the range of hyper-planes that perform similarly well, even if they have a radically different plane normal. This brings into question their utility for discovering new expression relationships in small or noisy data-sets, as it becomes difficult to distinguish between results that are significant, where all 'good' hyper-planes have very similar normal vectors, and those that are correspond to the 'best' but uninformative solution, where a wide range of normal vectors would perform nearly as well. In this chapter, we apply a Bayesian approach to training that is able to address this.

In Section 4.1.1, we discussed how SVMs can be represented as a sum of basis functions (Equation 4-5). The general class of models that take on this form are called Generalized Linear Models (GLMs) (Nelder and McCulagh 1983). During training, the selection of the weights is just a scaling factor for each subspace, stretching the dimensions that increase the accuracy of the model, and shrinking those that are irrelevant. From this point of view, the basis functions each define a dimension in the feature space under consideration.

Some of the basis functions will be highly correlated with one another. This means that using more than one of these will contribute little or no additional information. Other basis functions may simply be uninformative to the problem in hand. By defining some measure of the information contributed to the model by a given basis function, and the additional complexity of including that function, it is possible to make a trade-off between the simplicity of a model and how well it fits the data.

Bayes Theorem states how the probability of simultaneously observing two events is related to the probability of observing one event in isolation and the probability of

observing the second event given that we already know that the first one has occurred. Let us consider the case of observing a model, m and data, d.

Equation 5-1 Bayes Theorem

$$p(m,d) = p(m \mid d)p(d) \\ = p(d \mid m)p(m)$$

This equivalence can be re-arranged to express one of the conditional probabilities in terms of the independent probabilities and the other conditional probability. It is in this form that Bayes Theorem is most often presented.

Equation 5-2 Rearrangement of Bayes Theorem

$$p(m \mid d) = \frac{p(d \mid m)p(m)}{p(d)}$$

The terms in this form all have names in Bayesian statistical analysis.

Equation 5-3 Bayes Theorem in Words

$$posterior = \frac{likelihood \cdot prior}{evidence}$$

In the case of models and data, the posterior is the probability of our model (and associated parameters) given the data. The likelihood is the probability of observing the data given our model. The prior is the degree of belief we have that the model is sensible. The evidence is the probability of observing our data given any possible model, which in practice means the sum or integral of the probability of observing the data over all possible values for all parameters of the model.

One method for training GLMs which makes use of Bayesian statistics is the Relevance Vector Machine (RVM) (Bishop and E. 2000; Tipping 2000). In the case

of RVMs, the prior is chosen in such a way that it favours models where many of the weight parameters have values near to zero. For a particular parameter set to have a high posterior probability, the prior "cost" of any non-zero weights must be balanced by an increased value of the likelihood. If a particular basis function does not contribute to the likelihood sufficiently, then a greater overall posterior can be achieved by setting its weight to zero. RVMs can be trained by selecting parameters that maximize the posterior (Tipping 2000), or by fitting a variational approximating distribution to the posterior (Bishop and E. 2000).

A pure Java implementation of the RVM method has been implemented (Down 2003). This implementation uses patterns similar to the BioJava SVM implementation (Section 4.1.2) to insulate the optimiser from the data. An interface `BasisFunction` is provided that has a single method that returns the value of the basis function for a Java object. There is also an interface `BasisSource` that represents an iterater over a set of basis functions. The known implementations of RVMs (Bishop and E. 2000; Tipping 2000; Down 2003) all have space and time costs that scale very badly with the number of basis functions being considered. The API for Down's method employs the 'small working set' heuristic to work around this. During training, many weights become sufficiently close to zero to be discarded within a very few cycles of optimisation. This is exploited by setting a high and low water-mark for the set of basis symbols being considered. Initially, basis symbols are obtained from the `BasisSource` until the high water-mark is reached. The optimiser then runs until it has discarded enough basis functions that the low water-mark is reached. At that point, basis functions are added until the high water-mark is again reached, all parameters are re-initialised and the optimisation is resumed. This process is continued until the `BasisSource` has no more basis functions available. At this point,

the optimiser runs until the model converges. This heuristic keeps the cost of training a model with increasing numbers of basis functions proportional to the total number of basis functions that must be considered, and some function of the working set size. In practice, this makes some problems tractable that would be otherwise intractable.

The RVM API of Down's implementation interacts with the BioJava APIs for SVMs. Where appropriate, interfaces for representing training data and models are reused. In addition, there is adaptor code that allows a kernel function and a set of training objects to be viewed as a `BasisSource` over the implied basis functions (see Equation 4-4). In practice, very few lines of code need be changed to switch between analysing a data set with SVM and RVM methods.

5.4 Micro-array Classification Using a Support Vector Machine Implemented as a Linear Kernel RVM

To investigate the behaviour of SVMs when applied to a hard expression analysis problem, we applied them to the dataset described above (Perou, Sorlie et al. 2000).

The BioJava implementation of SVMs was used to construct a classifying support vector machine using the dot product (linear) kernel function to evaluate expression data. The kernel function was implemented so that the expression data was represented as an array of the log of the ratio between background and experimental sample levels. This was trained using the complete set of expression data described in Section 5.1 using the SMO training algorithm. The resulting model contained nearly all micro-arrays as support vectors. This suggests that the model was effectively memorizing the training set. We therefore decided not to further investigate the use of classically trained SVMs for this task, as they seem to be unable to model this problem.

To investigate whether the SVMs were extracting any significant data from the expression data-set, or just memorizing it as suspected, we applied an RVM approach (Section 5.3). An RVM was constructed with a `BasisSource` using the above training data and kernel function to generate basis functions (See Section 4.1.1, and in particular Equation 4-4 and Equation 4-5). The RVM was then trained using the complete set of micro-arrays. Given these basis functions, the RVM becomes equivalent to a Bayesian interpretation of the SVM. This RVM rejected all basis functions during training. This indicates that none of the SVM solutions using a linear kernel function robustly describes how to separate the pre- and post-treatment samples.

This negative result does not necessarily mean that this task could not be performed with either an SVM or an RVM using linear kernel functions, but that there was insufficient training data to support any parameters. By working with larger training sets, or more complex kernels, it may be possible to apply a kernel RVM to this data. However, this result does indicate one of the main benefits of RVM training over SVMs in that the RVM was able to indicate that no reasonable model could be produced. The SVM produced the best model that it could, which was of poor quality, but without any indication to the user that this was the case. Any predictions made on the basis of genes contributing to the separating hyper-plane are likely to have been incorrect, but there would have been no way to know this purely from the SVM results themselves.

5.4.1 Framework for Generalised-Linear-Models amenable to Expression Arrays

Given the failure of the linear kernel model description used above to discover expression differences resulting from treatment using doxorubicin, we now present an alternative way to model the problem.

An individual array measurement can be considered as a tuple of measurements with one dimension per spot on the array. This is a convenient interpretation for database storage and cluster analysis. Another point of view considers each spot to be the result of evaluating a probabilistic function on the particular sample (the log of the ratio of measured expression levels in experimental and background samples). This interpretation takes into account that the expression level measured is subject to noise. It transforms individual measurements (and by extension the individual genes) into entities amenable to hypothesis-directed reasoning using the RVM framework (Section 4.1.1), as now each measurement for each gene can be treated as the value of a basis function.

Consider a set of genes, G, a set of micro-arrays, A, and the function that retrieves the level for a gene on an array, $l(g \in G, a \in A)$. For any particular fixed g, there exists a conditioned version of this function, which we shall call $l^g(a)$. A GLM can then be constructed where the set of functions being evaluated is the set $l^g(.)$ for each gene. This model produces an output based upon a weighted sum of the log ratios of expression levels of multiple genes that is potentially predictive of some process.

Given any pre-defined classes by which the array measurements can be classified, a GLM can be estimated to perform that classification. If the sparse training approach is taken, then the hope is that the model will tend to extract key genes that have a type of

response that helps in the classification task, and will tend to discard all uninformative genes. This has the beneficial property of giving back a list of genes that are representative of each distinct response to the stimulus that aids in the classification task. If multiple genes share the same or similar expression profiles, the sparsity properties of the trainer will tend to find the statistically most representative member of that group and discard all others. For some uses of the method, such as where a complete list of significant genes would be useful (including those contributing similar information), this property is a disadvantage. In these cases, some further analysis of the data will be required to recover these other genes from the training data.

5.4.2 RVM Analysis Using the Small Working Set Heuristic

To evaluate this approach, a training set was constructed containing all of the before and after treatment measurements introduced above. The aim was to classify micro-arrays into those before and after doxorubicin treatment. An output of 1.0 would indicate that the method was certain that it was an example of 'before'. An output of 0.0 would indicate that the method was certain that it was an example of 'after'. A value between these two values indicates the degree of confidence that the sample belongs to one class or the other.

The number of basis functions to be evaluated was very large (one for each of the 8,102 genes). It was not practical to train the RVM with all of these simultaneously, so the small working set heuristic, described above (Section 5.3), was employed. The high water-mark was set to 90, and the low water-mark was set to 75. As long as the total number of basis functions needed for the task is below the low water mark, we could expect the result to be unaffected.

To check whether the heuristic altered the result the training was performed three times, using different permutations of the training data and of the order that the functions were added. The three models produced were identical (the same genes with weights within the bounds of numerical precision), and gave the model shown in Table 5-1. This suggests that, with this data set, the small working set heuristic works.

Table 5-1 GLM for all before-after pairs (to 4 s.f.)

Accession	Weight	Gene Name	Description
AA017544	-3.269	RGS1	Regulator of G-protein signalling 1
T72398	4.982	TDO2	Tryptophan 2,3-dioxygenase
AA040944	-6.299	FOS	Transcription factor involved in the apoptotic pathway

This model correctly classifies all of the training examples using the log-ratios of just three genes. Of course, training and testing on the same data-set is not robust for assessing how well models generalise, but the simplicity of the model suggests that this approach may work. Additionally, one of the three genes used is FOS, which is known to be involved in the apoptotic pathway activated in response to doxorubicin treatment (see Section 5.2).

To assess how reproducible these results where, we performed a “leave one out” cross-validation. For each of the forty micro-arrays, a prediction was made using a classifier trained on the remaining thirty-nine micro-arrays. The accuracy rate of the model for unseen data can then be estimated as the average accuracy of these forty predictions.

Of the 40 different models generated, 29 predicted the unseen item correctly. This is an accuracy rate of 72.5%, compared to the expected rate of 50%. All of the correct predictions typically had extreme probabilities (< 0.2 or > 0.8) where as the incorrect predictions were all relatively close to 0.5 (> 0.3 and < 0.7). 15 models used three genes, 23 used four genes and 2 used five genes. Across these models, a total of 22 different genes were used. Every model contained AA040944 (FOS). 22 of them used AA027832 (HBA2) and 17 used AA017544 (RGS1). These results are summarised in Table 5-2.

In the forty models generated by cross-validation, several of them use one of two alternative probes for the gene TOP2A. The degree of reproducibility or otherwise of the levels associated with those two probes can be taken as an indication of the quality of the data-set. Figure 5-1 shows a scatter plot with one data point for each of the 40 micro-arrays, and x, y co-ordinates given by the level of expression for the two TOP2A probes in a given micro-array. The levels have an R^2 value of 0.68, indicating that although they are correlated, there is a considerable degree of independent variation.

A summary of the expression data for these probes is displayed in Figure 5-2. As is seen from the graph, none of the probes used in the models have clearly separate distributions before and after treatment. FOS, which is used by every model generated during cross-validation, shows differences between the two groups, as does JUN, and to a lesser extent, both of the TOP2A probes. However, it should be clear from this that there is no one unambiguous indicator gene.

Given that the cross-validation procedure produced a range of different basis functions with a range of weights, it is interesting to consider what linear models can

be generated by combining these basis functions and their weights. This should give us some further indication of how important particular basis functions are.

One linear model can be obtained by taking the average weighting of each probe across all of the cross-validation models in which it takes part. The result of applying this to the micro-arrays is presented in Figure 5-3 as the scores produced prior to conversion to probabilities. This model misclassifies only four micro-arrays, giving a 90% accuracy rate.

This model does not take into account that some probes are present in fewer models. It is possible to reflect this by averaging the weights across all models, using a weight of zero where the probe is not used in a particular model. The result of applying this to the micro-arrays is presented in Figure 5-4. This model correctly classifies all of the micro-arrays. However, the associated confidences are lower, as demonstrated by the reduced magnitude of the outputs. The increase in accuracy of this model supports the idea that basis functions which are frequently present in different models are more informative to the classification task.

Using the contribution of just FOS to the model in Figure 5-4 (FOS level multiplied by its weight), all of the samples taken after treatment can be correctly identified, but 11 out of the 20 samples taken before treatment are misclassified. Similarly, using just the contribution of TDO2, all of samples taken after treatment can be correctly identified, but 4 of the samples taken before treatment are incorrectly predicted as being after treatment. This contrasts strongly with the behaviour of the contribution of the third component RGS1, which uniformly predicts all samples as belong to the before treatment class, with just one before and one after treatment sample predicted as after treatment. Each of the models generated during the cross validation procedure

contains exactly one probe that uniformly predicts all microarrays as belonging to one class (data not shown). We propose that the RVM is using these uniform predictors as a calibrated model of the level and variation inherent within this data set.

The aim of this RVM approach is to classify microarrays into two classes using the expression levels associated with each gene within each microarray. This methodology produces models that can be readily interpreted in terms of the contribution of each gene. However, it is not the primary aim of this method to indicate discriminating genes. A student t-test is more appropriate as a means for identifying genes with differential expression levels. This test calculates the probability that two sets of numbers have normal distributions that are distinguishable from one-another.

The student t-test scores associated with the range of levels in the before and after treatment groups is presented in Table 3-2. The column labelled TP contains the t-test scores for the two sets of microarrays taking into account the pairing between samples taken from the same patient before and after treatment. This information was not available to the RVM, and the student t-test scores assuming no such pairing are contained within the column labelled TS.

The probes for FOS and JUN have values that are extremely significant, indicating that there is very strong support for the hypothesis that the microarray levels before and after treatment come from different distributions. Generally, the t-test scores (both TS and TP) do not show any clear trend related to the rank of the probe in the table, or with the use of the probe as a uniform predictor. Although many of the probes used in the cross-validation models do have significant t-test scores, some do not, both at the 5 % and the 1 % significance level. Interestingly, many of the TP

scores are actually worse than the associated TS scores. It would be expected that in a system with low noise, the extra information provided by the sample pairing would lead to systematically greater significance. The presence of counter-examples may indicate that when considering individual genes, the level of noise in this data in some cases obscures the signal provided by the before and after treatment pairing.

Table 5-2 Genes used by cross-validation models

All information taken from the data files providing the expression data. Accession values of (*) indicate that the spot had no associated probe. TS is the value of the student t-test assuming the before and after samples to be unpaired. TP is the value of the student t-test taking into account that before and after samples are paired.

Probe	Accession	Symbol	Uses	TS (%)	TP (%)	Description
9016	AA040944	FOS	40	0.00	0.00	v-fos FBJ murine osteosarcoma viral oncogene homolog
8530	AA027832	HBA2	22	4.89	2.95	Hemoglobin, alpha 2
243	AA017544	RGS1	17	1.11	1.38	Regulator of G-protein signalling 1
2114	AA454668	PTGS1	11	0.24	0.07	Prostaglandin-endoperoxide synthase 1 (prostaglandin G/H synthase and cyclooxygenase)
6333	N50845		11	5.74	9.64	
5635	*		8	0.60	0.39	
6077	AA425316	LOC51700	7	1.78	2.02	Cytochrome b5 reductase b5R.2
7399	AA026682	TOP2A	5	0.71	0.60	Topoisomerase (DNA) II alpha (170kD)
3903	*		3	5.13	4.39	
3901	*		2	0.36	0.45	
5284	T72398	TDO2	2	7.33	0.83	Tryptophan 2,3-dioxygenase
6223	*		2	15.48	16.19	
6494	W96134	JUN	2	0.00	0.00	v-jun avian sarcoma virus 17 oncogene homolog
7956	T63045	IGL@	2	16.01	1.40	Immunoglobulin lambda locus
244	AA074224	RCV1	1	9.98	12.11	Recoverin
2753	*		1	4.63	6.26	
4468	AA453345	JAK2	1	6.99	3.93	Janus kinase 2 (a protein tyrosine kinase)
5002	AA620359		1	1.20	0.62	
6043	H87471	KYNU	1	20.95	4.81	Kynureninase (L-kynurenine hydrolase)
7704	N71028		1	0.99	0.17	
8494	*		1	4.58	8.33	
8719	AA504348	TOP2A	1	2.34	2.39	Topoisomerase (DNA) II alpha (170kD)

Error! Not a valid link.

Figure 5-1 Scatter Plot of the Two Topoisomerase II Probes Used.

There is one point for each of the 40 micro-arrays. The x values are the levels of the probe for AA504348, and the y values are the levels of the probe AA026682. Both of these are probes for the TOP2A gene. The R^2 value is the correlation between the levels measured for these two probes under identical conditions.

Error! Not a valid link.

Figure 5-2 Expression Levels for Each Probe Used

For each probe, there are three bars. Each data-point displays the mean level for a probe across a range of micro-arrays. The error bars display two standard deviations around the mean. In each case, the left-most bar corresponds to the mean and standard deviation of the probes level across the 20 micro-arrays taken before treatment, and the right-most par corresponds to the mean and standard deviations for the probe across the 20 micro-array measurements after drug treatment. Each data-point is labeled with the gene name if present. If this was not present, the accession number is used. If this was not available, the probe number is used. The probes are in the same order as Table 5-2.

Error! Not a valid link.

Figure 5-3 Average Weighs Across Relevant Models.

The samples after treatment are to the left, and samples before treatment are to the right. All prediction values are in the units of the GLM before conversion into probabilities. Values below 0 will map to probabilities below 0.5, and values above 0 will map to probabilities above 0.5. All of the before samples have been correctly classified. Four of the after samples are misclassified, and are indicated with an asterisk (*).

After Before

Error! Not a valid link.

Figure 5-4 Average Weights Across All Models

The samples after treatment are to the left, and samples before treatment are to the right. All prediction values are in the units of the GLM before conversion into probabilities. Values below 0 will map to probabilities below 0.5, and values above 0 will map to probabilities above 0.5. All of the samples have been correctly classified.

After Before

5.4.3 Function of Genes Identified by GLM Models

If the model learned a biologically significant signal, this should be reflected in the probes used to construct the model (as listed in Table 5-2). For many of the genes, this is indeed the case. Several genes known to be involved in the action of doxorubicin are present.

TOP2A directly interacts with doxorubicin, leading to the single-strand break mechanism of drug activity. This appears to be down-regulated in the group after treatment. This could be evidence that the cancers are developing doxorubicin resistance by repressing the TOP2A gene. Alternatively, potentially irreversible interactions between TOP2A and doxorubicin intercalated with DNA, or the relative lack of super-coiling due to many single-strand breaks may be fooling the regulatory mechanisms for TOP2A into behaving as if there are sufficient levels of the protein, leading to down-regulation of the gene.

JUN and FOS are part of the pathway that mediates apoptosis in response to excessive rates of single-strand breakages. Both of these appear to be up-regulated in the group after treatment. JUN and FOS form a transcription regulatory complex, and in cells responding to single-strand break stress, this complex interacts with the genes responsible for activating the apoptosis response. The resulting reduction in level of free JUN and FOS may cause their synthesis to be up-regulated to compensate.

RGS1 is a repressor of the G protein signalling that is involved in the regulation of b-cell activation and proliferation, as well being indicated in a range of cancers[35]. It appears to be marginally down-regulated after treatment. G proteins are involved in a

[35] See http://caroll.vjf.cnrs.fr/cancergene/CG516.html for a description of RGS1

wide range of signalling activities, and initiate MAP-kinase cascades. By down-regulating the repressor, the activity of the G proteins would be enhanced, increasing the strength of the signalling pathway. The single strand breaks introduced by Doxorubicin activity tend to arrest cell division. An increase in proliferation signals mediated by G protein signalling may compensate for this effect.

Two enzymes, TDO2 and KYNU, are present from the tryptophan metabolism pathway. Both of these enzymes appear to be down regulated in response to doxorubicin treatment. TDO2 catalysis the conversion of tryptophan to N-formyl-kynurenine. KYNU catalyses the conversion of this compound to formyl-anthranilate. It is intriguing that the models identified these two enzymes, given their proximity in a pathway. Intracellular levels of tryptophan around tumours have been shown to be abnormal (Iwagaki, Hizuta et al. 1995; Huang, Fuchs et al. 2002), but there is no clear indication of why this pathway should be important in response to doxorubicin.

The two genes HBA2 (haemoglobin alpha 2 subunit) and LOC51700 (cytochrome-b5 reductase) are present in several models. Both of these appear to be down regulated in response to doxorubicin treatment. Cytochrome-b5 and haemoglobin both require haem[36] for their production. Cytochrome-b5 reductase decreases the levels of available cytochrome-b5. Reduction in the levels of this enzyme would lead to increased levels of cytochrome-b5. A down regulation of HBA2 production would reduce the amount of haem becoming incorporated into haemoglobin. If both proteins are expressed within the same cells, these two processes would act together to increase the level of cytochrome-b5.

---

[36] See http://www.genome.ad.jp/dbget-bin/www_bget?compound+C00032 and links from that page for a more full description of haem and the Prophyrin metabolism pathway

As the samples used for microarray analysis were obtained from biopsies, it is inevitable that they represent expression levels from a range of different cell types. It is possible that within this population there were immature red blood cells. Although the nucleated red blood cell precursors are present only in the bone-marrow, there is a stage in their differentiation intermediate between this and mature red blood cells that contains mitochondria and messenger RNA. For a couple of days, these are present in the blood stream (Gilbert 2003). During this maturation stage, haemoglobin is synthesized. It is possible that the chemotherapy results in a decrease rate of red blood cell production. This would lead to a decreased number of maturing red blood cells in the circulatory system, and therefore a lower measured level of the haemoglobin mRNA.

5.5 Conclusions, Applications and Future Work

In this chapter, we have shown that RVMs can be used in the analysis of expression data that contains few samples and is noisy. The RVM was both able to perform the required classification task, and the model produced has clearly identified biologically relevant genes.

Cluster analysis of this expression data does not help in discovering genes that have modified expression levels in response to treatment with doxorubicin. Clustering by correlation co-efficient identifies clusters containing pairs of micro-arrays from a single patient. It does not produce clusters corresponding to all micro-arrays pre- or post-treatment, indicating that the history of the cell line is the primary signal in the expression profiles.

When an SVM was trained using the tumour sample expression data, it appeared to memorize the training set. When the same model was trained using the probabilistic RVM trainer, the RVM rejected the hypothesis that the data was separable using the

linear kernel function. This indicated that with just 20 samples, a conventional SVM could not be constructed to classify these samples into pre- and post-treatment.

Using an alternative strategy, an RVM was constructed with one basis function for each unique probe used to measure the level of a gene, and applied to learn a discriminator to predict whether tumour samples were pre- or post-treatment. It was able to learn signals that correlated with the treatment status. The function learned by RVM did appear to display all of the expected traits of sparsity, simplicity and generalization expected from this training method. Of course, given 8,102 genes to choose from, a model could trivially be constructed that performed very well on the training set. However, this would not be expected to generalize to unseen data. The results of the cross-validated training indicate that the models do generalize regardless of the sub-set used for training and testing, and that the models do not purely contain some statistically aberrant signal present by chance in the training set. With larger training sets, it should be possible to learn models with better estimates for which genes are informative and are less prone to over-fitting.

Some of the probes identified as basis functions during cross-validation appear to show differences in their average levels before and after drug treatment. Others do not. However, the RVMs are not looking at genes in isolation, but rather looking at interactions between them. A number of genes were identified as indicators that make clear biological sense, given the known action of doxorubicin (JUN, FOS, topoisomerase II). A number of others are not surprising, such as those associated with signalling cascades. Others, such as TDO2 and KYNU are implicated by their presence in the model and their differences in mean level before and after treatment.

A final group appear to be used by the model as a measure of the noise in the system, or to obtain a baseline from which all other levels can be calculated.

This use of RVMs is potentially applicable to any situation where large numbers of expression levels have been measured and a test is required which will indicate which of these are informative for a particular biological response. The classifier generated can be used to classify new data. RVMs can be trained using any set of basis functions. This is applicable to a wide range of situations, for example, screening expression levels from patient samples to estimate which of a range of anti-cancer drugs may be an appropriate treatment. It is possible to combine expression data with any other measurements. For example, expression levels could be combined with information about the presence or absence of SNPs, direct biological measurements, such as pulse or breathing rate, and so forth into a single predictive model.

The RVMs have two advantages over support vector machines (SVM) for this type of data. Firstly, to evaluate an SVM it is necessary to calculate products for every gene on the micro-array. The SVM will be invalid if all of the genes on the original micro-array are not also measured in the clinical sample from the patient, as the full dot product between the support vectors and the sample cannot be calculated. In addition, this requires one multiplication per micro-array spot per support vector. An RVM of the form described above only requires that the genes that are used by the model be within the set measured in the patient sample. It requires one multiplication for each gene that is used as a basis function to the learned weight.

Secondly, the GLMs produced by RVMs give an indication of confidence in their prediction. This potentially allows the person interpreting the model output to make sensible judgements about how to use the model's prediction (for example, ignoring

predictions with very low confidence). SVMs give an output value, but the absolute scale of this number is dependant on the distance from the separating hyper-plane of the two support vectors corresponding to the closest correctly classified data points from each class. Two SVMs using different support vectors will have incomparable scales, making it impossible to compare these values directly.

To evaluate the effectiveness of these RVMs for other classification tasks using micro-array data, more data sets need to be analysed. The data set used here was both small and noisy. It is to be expected that with larger data sets containing cleaner expression levels, that much higher levels of classification accuracy can be achieved. In the case where this method is used as a way of identifying biologically relevant genes, methods need to be developed to extract models with more genes. This could be achieved by training the model repeatedly, removing all probes identified by previous models until the model does not perform the classification task. Alternatively, it may be possible to look at the information each indicator gene is contributing, and to use some form of hierarchical or single-linkage clustering to identify those with patterns of expression that share information with it. The RVM could be modified to indicate if each basis was rejected because it did not contribute to the accuracy of the model, or because it duplicated information present in another basis.

Since this work was carried out, related relevance-based approaches have begun to emerge, for example (Li, Campbell et al. 2002) and Gene-Rave[37]. However, these

[37] See http://www.bioinformatics.csiro.au/GeneRave/products.html and the examples link from this page

methods do not seem to be producing results with quite the same high level of sparsity our method generates. Neither of these methods has as yet solved the question of how to retrieve the indicator genes removed from the model because they give information correlated to that of the selected relevant genes.

Concluding Remarks

The vast volumes of biological data being produced now overwhelm the traditional paradigm of individual scientists studying individual results and making and testing individual hypotheses. In this dissertation, I present tools and methods that allow data sets of genomic scales to be explored, analyzed and learned from. The BioJava project provides programming tools for manipulating genomic data sets. HMMs can be constructed which leverage un-supervised learning techniques to elucidate the inherent structure of chromosomes. SVMs and latterly RVMs can be used to perform regression and classification tasks on large quantities data with an unprecedented degree of sparsity and generalization. Here, they are used for the diverse tasks of predicting recombination rates and classifying tumour samples into those treated and un-treated with Doxorubicin.

During my PhD studies, I have used BioJava and its machine learning implementations in a range of other situations, which are not discussed in detail here. This was partly to define the limitations of the methods and partly for scientific exploration. Briefly, SVMs were applied to a wide range of regression and classification tasks. These included the implementation of an e-mail spam filter, assessing the accuracy of gene predictions given the outputs of multiple programs and curve smoothing for recombination rates. RVMs were applied to an equally wide range of problems, including predicting protein secondary structure elements, sequence comparison using HMM kernels and simultaneous estimation of expression profile class and promoter structures. HMMs were used to model 3-D DNA structure using a multinomial Gaussian emission state, Gibbs-sampling of expression profiles, promoter finding, protein secondary structure prediction by pair wise alignment and

HMM-based kernel functions. This is by no means an exhaustive list of mini-projects undertaken within the past four years, but gives a flavour of the range of problems that can be tackled using these technologies.

Since this thesis was written, BioJava has continued to develop (as discussed in chapter 2), demonstrating that the original APIs are both flexible and sufficient, allowing a wide degree of reuse and extension.

None of the methods presented here are limited to the problems to which they were applied. The task ahead is to use these and other technologies to make new discoveries about how genomes are structured, function, evolve and fail. The possible applications are wide-ranging; medicine, agriculture, bio-engineering, palaeontology, to name but a few. I look forward to seeing where this leads us.

Forgive us for what we have done and what we have left undone

(Extract from the Anglican Order of Service)