Sparse Bayesian Promoter Based Gene Classification

Kee Khoon Lee[†], Gavin C. Cawley[†] and Michael W. Bevan[‡]

[†]School of Computing Sciences
University of East Anglia
Norwich, U.K. NR4 7TJ
{kkl,gcc}@cmp.uea.ac.uk

[‡]John Iness Institute,
Norwich Research Park,
Norwich, U.K. NR4 7UH,
michael.bevan@bbsrc.ac.uk

Abstract. A method to distinguish between co-regulated genes that are up- or down-regulated under a given treatment, based on the composition of the upstream promoter region, would be a valuable tool in deciphering gene regulatory networks. Ideally, the classification should be based on a small number of regulatory motifs, whose presence in the promoter region of a gene induce a significant effect on its transcriptional regulation. In this paper, we investigate the use of Relevance Vector Machines for this task, and present initial results of an analysis of glucose response in the model plant *Arabidopsis thaliana*, that has revealed novel biological information.

1 Introduction

The interpretation of DNA sequence and hence the ability to use the sequence to understand biological processes such as growth, development and disease resistance is one the major challenges in biology. To address these needs a catalog or "parts list" of the structural and functional components encoded in DNA sequence needs to be assembled. The information content of DNA is transcribed into RNA for further processing in cells, either as a structural RNA or into protein sequences through the well- known triplet codons. Most higher organisms, from flies to plants to humans, have about 12,000–30,000 genes that encode different proteins, the building blocks of cells. The information specifying the time, cellular location and amount of RNA transcribed from each of these genes is specified by extensive DNA sequences adjacent to the genes. Most biological processes are regulated by coordinating the transcription of multiple genes, therefore assessing and understanding the sequence information directing transcription is of fundamental and widespread interest.

This work was supported by a grant from the Biotechnology and Biological Sciences Research Council (BBSRC) of the United Kingdom, grant numbers 83/EGM16126 (to M.W.B.) 83/EGM16128 (to G.C.C.).

The transcription of genes is regulated by proteins called transcription factors that interact with specific DNA sequences in regulatory regions of genes. Combinations of different transcription factors binding to regulatory regions provide the high specificity of gene expression. By identifying these regulatory sequences and their relative positions in genes it will be possible to establish the complex regulatory circuitry coordinating the expression of thousands of genes necessary to execute a biological process.

The transcription of all genes can now be accurately measured using microarray technology in many species. By establishing relationships and dependencies between transcript abundance and regulatory sequences it may be possible to identify specific combinations of transcription factor binding sites that confer transcript levels. We propose the use of the Relevance Vector Machine (RVM) [6] to classify co-regulated genes as a means of identifying putative transcription factor binding sites. As an experimental system we use microarray and genome data from the plant *Arabidopsis*, which is completely sequenced and has a well characterised and compact genome. RVM classification of gene expression in response to the simple nutrient glucose identified a large number of putative transcriptional regulatory circuits that were verified by subsequent experiments.

2 The Relevance Vector Machine

In a pattern recognition setting, the Relevance Vector Machine (RVM) [6] can be viewed as a simple logistic regression model, with a Bayesian Automatic Relevance Determination (ARD) prior [3] over the weights associated with each feature in order to achieve a parsimonious model. Let $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ represent an alphabet of symbols representing the bases adenine, cytosine, guanine and thymine respectively. The RVM constructs a decision rule from a set of labelled training data,

$$\mathcal{D} = \{(\boldsymbol{x}_i, t_i)\}_{i=1}^{\ell}, \qquad \boldsymbol{x}_i \in \mathcal{A}^{n_i}, \qquad t_i \in \{0, \ +1\},$$

where the input patterns, $\boldsymbol{x}_i$, consist of strings drawn from $\mathcal{A}$ of varying lengths, describing the upstream promoter regions of a set of co-regulated genes. The target patterns, t_i, indicate whether the corresponding gene is up-regulated (class $\mathcal{C}_1$, $y_i = +1$) or down-regulated (class $\mathcal{C}_2$, $y_i = 0$) under a given treatment. The RVM constructs a logistic regression model based on a set of sequence features derived from the input patterns, i.e.

$$p(\mathcal{C}_1|\boldsymbol{x}) \approx \sigma\left\{y(\boldsymbol{x}; \boldsymbol{w})\right\} \qquad \text{where} \qquad y(\boldsymbol{x}; \boldsymbol{w}) = \sum_{i=1}^{N} w_i \varphi_i(\boldsymbol{x}) + w_0, \tag{1}$$

and $\sigma\left\{y\right\} = (1 + \exp\{y\})^{-1}$ is the logistic inverse link function. In this study a feature, $\varphi_i(\boldsymbol{x})$, represents the number of times an arbitrary substring, $s_i \in \mathcal{A}^d$, ocurrs in a promoter sequence $\boldsymbol{x}$. A sufficiently large set of features is used such that it is reasonable to expect that some of these features will represent oligonucleotides forming a relevant promoter protein binding site and so provide discriminatory information for the pattern recognition task at hand. Assuming a Bernoulii distribution for $P(t|\boldsymbol{x})$, the *likelihood* of the training data, $\mathcal{D}$, can be written as

$$P(\mathcal{D}|\boldsymbol{w}) = \prod_{i=1}^{\ell} \sigma\left\{y(\boldsymbol{x}_i;\boldsymbol{w})\right\}^{t_i} \left[1 - \sigma\left\{y(\boldsymbol{x}_i;\boldsymbol{w})\right\}\right]^{1-t_i} \tag{2}$$

To form a Bayesian training criterion, we must also impose a prior distribution over the vector of model parameters or *weights*, $p(\boldsymbol{w})$. The RVM adopts a separable Gaussian prior, with a distinct hyper-parameter, α_i, for each weight,

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{N} \mathcal{N}(w_i|0, \alpha_i^{-1}). \tag{3}$$

The optimal parameters of the model are then given by the minimiser of the penalised negative log-likelihood,

$$\log\left\{P(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha})\right\} = \sum_{i=1}^{\ell}\left[t_i \log y_i + (1-t_i)\log(1-y_i)\right] - \frac{1}{2}\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}. \tag{4}$$

where $y_i = \sigma\left\{y(\boldsymbol{x}_i;\boldsymbol{w})\right\}$ and $\boldsymbol{A} = \mathrm{diag}(\boldsymbol{\alpha})$ is a diagonal matrix with non-zero elements given by the vector of hyper-parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$. This is achieved via the efficient Iteratively Re-Weighted Least Squares (IRWLS) algorithm [4]. Next, Laplace's method is used to obtain a Gaussian approximation to the posterior density of the weights,

$$p(\boldsymbol{w}|\mathcal{D},\boldsymbol{\alpha}) \approx \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu},\boldsymbol{\Sigma}), \tag{5}$$

where the posterior mean and covariance are given by

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{B}\boldsymbol{t}, \qquad \text{and} \qquad \boldsymbol{\Sigma} = \left[\boldsymbol{\Phi}^T\boldsymbol{B}\boldsymbol{\Phi} + \boldsymbol{A}\right]^{-1}$$

respectively, $\boldsymbol{\Phi}$ is an $\ell \times N$ matrix of features for each promoter in the training set and $\boldsymbol{B}$ is a diagonal matrix with non-zero elements $b_{ii} = y_i(1-y_i)$. The hyper-parameters are then updated in order to maximise their marginal likelihood, $p(\mathcal{D}|\boldsymbol{\alpha})$, according to the efficient update formula

$$\alpha_i^{\text{new}} = \frac{\gamma_i}{\mu_i^2} \qquad \text{where} \qquad \gamma_i = 1 - \alpha_i\Sigma_{ii}. \tag{6}$$

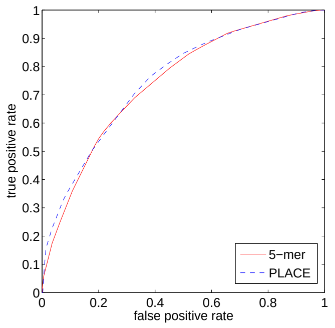
This process is repeated until an appropriate convergence criterion is met (see [6] for details). The maximisation of the marginal likelihood, or *evidence*, for the hyper-parameters, $\boldsymbol{\alpha}$, leads to the hyper-parameters associated with uninformative features becoming very large. This in turn forces the value of the associated weight essentially to zero, allowing redundant features to be easily identified and pruned from the model. Given a sufficiently rich set of sequence features, it seems reasonable to suggest that the features retained by the RVM *may* represent (parts of) transcription factor binding sites as they provide discriminatory information distinguishing between up- and down-regulated genes.

3 Results

Initial experiments investigate the response of *Arabidopsis* to glucose. Seedlings were grown in liquid culture for 7 days on low sugar concentrations (0.5% glucose) and constant light to abrogate light responses. After 7 days growth, the medium was replaced by a glucose fee medium for 24 hours and then glucose or mannitol (a non-toxic, non-metabolised sugar acting as an osmotic control) were added to 3% w/v. Treatments were designed to reveal transitions in gene expression from a sugar-restricted to a sugar-replete state. Microarray ananlysis, using Affymetrix ATH1 GeneChips, indicated that a set of 1,844 genes demonstrated a change in expression of 2.5-fold or more at the 2, 4 and 6 hour time points, of these 1,051 were found to be up-regulated and 793 down-regulated. Full details of the experimental method are given in Li *et al.* [2]. The learning task for the RVM was then to select discriminative features characterising the promoters of these co-regulated genes, allowing up-regulated genes to be distinguished from the down-regulated genes.

A database containing approximately 1,000 base pairs of DNA sequence data upstream from the initial `ATG` codon of each of the co-regulated genes was assembled. Two sources of candidate transcription factor binding sites were investigated. The first approach simply selects all 1024 possible strings of length five drawn from $\mathcal{A}$. Oligonucleotides of less than five bases will ocurr so frequently in all promoters as to be unlikely to provide useful discriminatory features. Features representing longer oligonucleotides will be more specific, and being only be found in the promoters of a few genes will again be unlikely to provide good discriminatory features, as their coverage is low. Features representing oligonucleotides of length five were found to be a good compromise as they are sufficiently long to adequately characterise the *core* of a transcription factor binding site, without being too specific or too pervasive. The transcription binding site can appear on either strand of the DNA double helix, and so *complementary* features, such as `ACTG` and `TGAC`, can be combined, leaving only 512 candidate features. The second approach utilises features formed from known transcription factor binding sites drawn from the PLACE database [1]. This is a repository compiled from published reports of motifs corresponding to known plant cis-acting DNA regulatory elements. A total of 253 such motifs were found to be present in the promoters of the genes found to be co-regulated in response to glucose. For computational expedience, features ocurring in fewer that 10 promoters were discarded due to insufficient coverage.

The performance statistics for RVM classifiers based on features drawn from the set of all 5-mers and from the PLACE database, presented in this section,



(a)

		Observed	
		T	**F**
Predicted	**T**	833	337
	F	218	456

(b)

		T	**F**
Predicted	**T**	794	325
	F	257	468

(c)

Figure 1: Receiver operating characteristic convex hull (ROCCH) curves (a) and confusion matrices for RVM classifiers based on PLACE (b) and 5-mer (c) candidate features.

were evaluated using a 10-fold cross-validation procedure [5]. Figure 1 (a) shows the convex hull of the receiver operating characteristic curves for RVM classifiers, the classifier based on PLACE features performs marginally better than that based on 5-mer features. This is also reflected in the area under the ROCCH statistic, having a value of 0.754 for the PLACE feature set and 0.745 for the 5-mer feature set. The RVM based on the PLACE feature set classified 69.9% of the patterns correctly, whereas the RVM based on 5-mer features classified only 68.4% correctly. The confusion matrices for both classifiers are given in Figure 1 (b) and (c).

Table 1 lists a sample of the strongest PLACE features retained by the RVM (see Li *et al.* [2] for additional features), showing a mix of expected elements (e.g. IBOX & AMMORESIIUDCRNIA1) as well as some less obvious choices (e.g. TELOBOXATEEF1AA1 & DRECRTCORENT). The TELOBOX and DRE elements had not previously been implicated in glucose-responsive gene expression, and so we experimentally verified the significance of these features. A further micorarray analysis was performed, involving plants engineered to contain a reporter plasmid with a minimal promoter comprised of tetramers of these regulatory elements. Both the TELOBOX and DRE elements were found to confer glucose-responsive expression. Reassuringly, many of the prominent 5-mer features chosen by the RVM represent (parts of) PLACE element features, such as `GGATA`, `GATAA` and `ACCCT`, corresponding to the MYBST1, IBOXCORE and TELOBOX elements.

4 Conclusions

In this paper we have demonstrated that the Relevance Vector Machine can be used for sparse Bayesian selection of putative transcription factor binding sites based on microarray gene expression results for co-regulated genes. Our initial results appear promising, already having predicted novel functions for two known transcription factors that have been verified experimentally. Further work will investigate modelling expression dynamics and will also include sequence features from exon, intron and 3' regions.

References

[1] K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Research*, 27(1):297–300, 1999.

[2] Y. Li, K. K. Lee, S. Hadingham, C. Smith, S. Walsh, K. Sorefan, Cawley G., and M. W. Bevan. Promoter classification using machine learning procedures reveals glucose and ABA-regulated transcription in *Arabidopsis*. *The Plant Cell* (submitted), 2004.

[3] D. J. C. MacKay. Bayesian methods for backpropagation networks. In *Models of Neural Networks III*, chapter 6, pages 211–254. Springer, 1994.

[4] I. T. Nabney. Efficient training of RBF networks for classification. In *Proceedings of the Ninth Int. Conf. on Artificial Neural Networks*, volume 1, pages 210–215, September 7–10 1999.

[5] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, B*, 36(1):111–147, 1974.

[6] M. E. Tipping. Sparse Bayesian learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

Table 1: Some of the PLACE features selected by the RVM in all cross-validation trials and their average weights and potential target genes.

Element ID	**Weight**	**Sequence**	**Target Genes**
TELOBOXATEEF1AA1	+2.990	AAACCCTAA	Ribosomal proteins, helicases, translation initiation factors
AMMORESIIUDCRNIA1	+1.335	GGWAGGGT	Carbon metabolosm and DNA replication proteins
QARBNEXTA	+1.103	AACGTGT	Phenylpropanoid synthesis and starch metabolism
BS1EGCCR	+0.964	ACGGGG	Anthocyanin and glucose metabolism enzymes
DRECRTCORENT	+0.808	RCCGAC	NaCl, cold & other stress related genes
IBOXCORENT	-3.320	GATAAGR	Light related proteins, T6PS,
IBOXCORE	-2.143	GATAA	enzymes, mitochondrial
IBOX	-0.711	GATAAG	biogenesis proteins
MYBST1	-3.214	GGATA	Light and protein degradation proteins