Predicting Sugar Regulation in *Arabidopsis thaliana* using Kernel Learning Methods

Kamel Saadi, Kee-Khoon Lee, Gavin C. Cawley
School of Computing Sciences
University of East Anglia
Norwich NR4 7TJ U.K.
E-mail: {ks,kkl,gcc}@cmp.uea.ac.uk

Michael W. Bevan
John Innes Centre
Norwich Research Park
Norwich NR4 7UA U.K.
E-mail: mike.bevan@bbsrc.ac.uk

***Abstract*— The ability to predict the transcriptional regulation of genes, based on the composition of the upstream promoter region, would be a useful step in deciphering gene regulatory networks in eukaryotic organisms. In this paper we perform optimally regularised kernel Fisher discriminant (ORKFD) analysis of the upstream promoter sequences of genes to predict whether they are up- or down-regulated in response to glucose in the model plant *Arabidopsis thaliana*. Three feature selection strategies are investigated, namely use of known promoter motifs drawn from the PLACE database, explicit enumeration of all possible k-mers and the use of the mismatch kernels (which effectively permits the construction of a linear model in the space of all possible k-mers with up to m mismatches). The leave-one-out cross-validation (LOOCV) error rate indicates that approximately two-thirds of of the observed regulatory behaviour can be inferred by the presence of particular motifs in the upstream promoter sequence. The analysis has yielded novel biological insight, which has since been confirmed experimentally *in vivo*.**

I. Introduction

The genomes of animals, plants and micro-organisms are comprised of thousands of genes whose expression is regulated to co-ordinate growth and development. One of the most exciting and complex challenges in biological research is to understand the mechanisms regulating gene expression, and to understand how gene expression is integrated in space and time. Multicellular organisms have a relatively complex gene structure comprising the coding regions or exons that encode protein sequence, introns that separate the individual exons comprising a gene, and conserved regulatory sequences flanking the gene that confer specific patterns of expression and direct the start and stop points for messenger RNA synthesis. Strategies to identify DNA sequence motifs implicated in regulating gene expression are less clear and effective, because these motifs are poorly defined, are relatively short, and are not generally strongly conserved within and between species. In this paper, we aim to estimate the degree to which up- or down-regulation can be inferred from the presence or absence of these conserved regulatory motifs, using glucose regulation in *A. thaliana* as a test-case. For this study we adopt kernel learning methods (see e.g. [1–3]), which facilitate the construction of classifiers acting directly on biological sequence data.

A. A Brief Overview Gene Regulation in Eukaryotes

The DNA of eukaryotic organisms is arranged in a number of chromosomes, each of which is a single molecule consisting of a linear polymer comprised of four different basic building blocks, known as “nucleotides” (adenine, cytosine, guanine and thyamine, usually represented by the letters `A`, `C`, `G` and `T` respectively). Each chromosome is divided into *genes*, each of which contains the genetic information specifying the sequence of amino acids forming a particular protein. Figure 1 shows a schematic representation of the structure of a gene in a eukaryotic organism. The DNA sequence of a gene is comprised of two sections, the transcribed region and the promoter region. For the synthesis of a protein to occur, a copy of the transcribed region must first be made in messenger RNA (mRNA). The transcribed region consists of *exons*, which specify the sequence of amino acids comprising the protein, separated by *introns*. Before leaving the cell nucleus, the mRNA is *spliced* to remove the sections corresponding to the introns. Some genes may be spliced in a number of alternative configurations, allowing a number of related proteins to be synthesised from a given gene. The exons consist of a sequence of *codons*, groups of three contiguous nucleotides, each of which specifies one of the twenty amino acids concatenated to form a protein.

The concentration of a protein within the cell body then depends on the rate at which the protein is synthesised and degraded by the biochemical machinery of the organism. The primary control on the rate of synthesis of a protein is provided by *transcriptional regulation*, which governs the rate at which mRNA copies of the coding region of the gene are transcribed. The *promoter* is a region of the DNA sequence that occurs “upstream” of the transcribed region of the gene. The transcription of the majority of eukaryotic genes is performed by an enzyme called RNA polymerase II, which moves downstream along the DNA sequence transcribing the mRNA copy one nucleotide at a time. In order for RNA polymerase II to bind to the appropriate starting position, a number of proteins known as *transcription factors* must first bind onto transcription factor binding sites within the promoter region. Transcription factors can act to encourage or inhibit transcription, in which case they are called *enhancers* or *silencers* respectively. Combinations

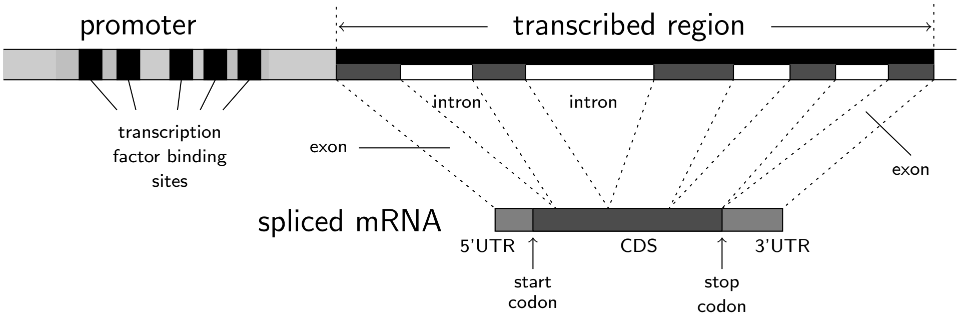

Fig. 1. Schematic representation of the structure of the eukaryotic gene, after Zien [4].

of different transcription factors binding to regulatory regions provide the high specificity of gene expression. Note that the sequence of bases forming a transcription factor binding site results in a specific conformation of (usually) the major groove of the double-helix structure, which matches the shape of part of the transcription factor. The complementary sequence (formed by reversing the order of bases and exchanging As and Ts and Cs and Gs) results in an identical conformation, but with the opposite orientation with respect to the transcribed region. The orientation of a transcription factor is not thought to be significant, and so the sequence corresponding a binding site and its complement are considered to be equivalent representations. The transcription of genes is then *regulated* by the nuclear concentrations of these transcription factors. For a more detailed, but accessible introduction to gene regulation in both prokaryotes and eukaryotes, see Alberts *et al.* [5].

B. Promoter-based Gene Classification

By identifying transcription factor binding sites and their relative positions in promoter region of genes it will be possible to establish the complex regulatory circuitry coordinating the expression of thousands of genes necessary to execute a given 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 kernel learning methods (e.g. [1–3]) to classify co-regulated genes, whose transcriptional abundance increases or decreases in response to a given environmental stimulus, 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. 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.

The use of kernel learning methods provides a flexible means to efficiently investigate ways in which to select discriminative features for classification. Here we investigate features selected from a database of known transcription factor binding sites, selection from the set of all possible $k-$mers and through the use of the spectrum and mismatch kernels, linear combinations of all possible $k-$mers (perhaps allowing mismatches). An efficient kernel learning algorithm, namely optimally regularised kernel Fisher discriminant (ORKFD) analysis, which provides a computationally efficient means of constructing a kernel machine with the regularisation tuned so as to minimise the leave-one-out cross-validation error. This ensures that the complexity of the model is well-matched to the complexity of the learning task.

II. Optimally Regularised Kernel Fisher Discriminant Analysis

In this section, we give a brief review of the optimally regularised kernel Fisher discriminant analysis algorithm introduced by Saadi *et al.* [6]. Assume we are given training data $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell\} = \{\mathcal{X}_1, \mathcal{X}_2\} \subset \mathbb{R}^d$, where $\mathcal{X}_1 = \{\boldsymbol{x}_1^1, \boldsymbol{x}_2^1, \ldots, \boldsymbol{x}_{\ell_1}^1\}$ is a set of patterns belonging to class $\mathcal{C}_1$ and similarly $\mathcal{X}_2 = \{\boldsymbol{x}_1^2, \boldsymbol{x}_2^2, \ldots, \boldsymbol{x}_{\ell_2}^2\}$ is a set of patterns belonging to class $\mathcal{C}_2$; Fisher's linear discriminant (e.g. [7, 8]) attempts to find a linear combination of input variables, $\boldsymbol{w} \cdot \boldsymbol{x}$, that maximises the average separation of the projections of points belonging to $\mathcal{C}_1$ and $\mathcal{C}_2$, whilst minimising the within class variance of the projections of those points. The innovation introduced by Mika *et al.* [9] is to construct Fisher's linear discriminant in a fixed feature space $\mathcal{F}$ $(\boldsymbol{\phi} : \mathcal{X} \to \mathcal{F})$ induced by a positive definite *Mercer* kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defining the inner product $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}')$. The kernel Fisher discriminant(KFD) is then given by the kernel expansion,

$$f(\boldsymbol{x}) = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b. \tag{1}$$

It is well known that Fisher discriminant analysis is equivalent to linear least-squares regression on the class labels (e.g. [8]), and so the optimal parameters $\boldsymbol{\alpha}$ and b are given by the solution of the following system of linear equations (Xu *et al.* [10]):

$$\begin{bmatrix} \boldsymbol{K}\boldsymbol{K} + \mu \boldsymbol{I} & \boldsymbol{K}\boldsymbol{1} \\ (\boldsymbol{K}\boldsymbol{1})^T & \ell \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{K} \\ \boldsymbol{1} \end{bmatrix} \boldsymbol{y}, \tag{2}$$

where $\boldsymbol{1}$ is a column vector of ℓ ones and $\boldsymbol{y}$ is a column vector with elements $y_i = \ell/\ell_j \ \forall i : \boldsymbol{x}_i \in \mathcal{X}_j$, $\boldsymbol{K} = [k_{ij} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^{\ell}$ is the kernel or Gram matrix and μ is a regularisation parameter [11] controlling the bias-variance trade-off [12]. The KFD classifier has been shown experimentally to demonstrate near state-of-the-art performance on a range of artificial and real world benchmark datasets [9] and so is worthy of consideration for small to medium scale applications, such as that considered here. The key step in maximising generalisation performance is model selection, i.e. the choice of good values for kernel and regularisation parameters. The leave-one-out cross-validation error rate gives an almost unbiased estimate of the probability of test error [13], and so provides an attractive model selection criterion. In the remainder of this section, we show that the regularisation parameter of a KFD classifier can be efficiently tuned so as to minimise the leave-one-out error with a computational cost of only $\mathcal{O}(\ell^2)$ operations, giving rise to the optimally regularised kernel Fisher discriminant (ORKFD) classifier.

A. Kernel Fisher Discriminant Analysis in Canonical Form

In this paper we present an efficient algorithm for approximate cross-validation of kernel Fisher discriminant models, providing a practical criterion for model selection. The system of linear equations (2) can be written more concisely in the form

$$\boldsymbol{\beta} = \left[\boldsymbol{Z}^T\boldsymbol{Z} + \boldsymbol{R}\right]^{-1} \boldsymbol{Z}^T\boldsymbol{y}, \qquad (3)$$

where $\boldsymbol{Z} = [\boldsymbol{K} \ \boldsymbol{1}]$ and $\boldsymbol{R}$ is a diagonal matrix with elements given by the vector of regularisation parameters $\boldsymbol{\mu}$. Let $\boldsymbol{V}$ be an orthogonal matrix, the columns of which are the eigenvectors of $\boldsymbol{Z}^T\boldsymbol{Z}$, and $\boldsymbol{\Lambda}$ be a diagonal matrix containing the corresponding eigenvalues $\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_\ell \geq 0$, such that $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$ and $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$. The principal components of $\boldsymbol{Z}$ are then given by the columns of $\boldsymbol{U} = \boldsymbol{Z}\boldsymbol{V}$; note that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{\Lambda}$. The system of linear equations (3) can then be expressed in *canonical form* [14] as

$$\boldsymbol{\alpha} = \boldsymbol{C}^{-1}\boldsymbol{U}^T\boldsymbol{y} = [\boldsymbol{\Lambda} + \boldsymbol{R}]^{-1}\,\boldsymbol{U}^T\boldsymbol{y}, \qquad (4)$$

where $\boldsymbol{\alpha} = \boldsymbol{V}^T\boldsymbol{\beta}$. The principal advantage of expressing the system of linear equations (3) in this form is that the matrix $\boldsymbol{C}$ is diagonal, and so can be inverted in linear time, i.e. $\mathcal{O}(\ell)$ operations, and the parameters of the KFD can be updated following a change in the vector of regularisation parameters with a computational complexity of only $\mathcal{O}(\ell)$ operations.

B. Efficient Leave-One-Out Cross-Validation

At each step of the leave-one-out cross-validation procedure, a kernel Fisher discriminant classifier is constructed excluding a single example from the training data. The vector of canonical model parameters, $\boldsymbol{\alpha}_{(i)}$ at the i^{th} step, in which pattern i is excluded, is then given by the solution of a modified system of linear equations,

$$\boldsymbol{\alpha}_{(i)} = \left[\boldsymbol{R} + \boldsymbol{U}_{(i)}^T\boldsymbol{U}_{(i)}\right]^{-1} \boldsymbol{U}_{(i)}^T\boldsymbol{y}$$

where $\boldsymbol{U}_{(i)}$ is the sub-matrix formed by omitting the i^{th} row of $\boldsymbol{U}$. Note that $\boldsymbol{U}_{(i)}^T\boldsymbol{U}_{(i)}$ is in general no longer diagonal, and so the most computationally expensive step is again the inversion of the matrix $\boldsymbol{C}_{(i)} = \left[\boldsymbol{R} + \boldsymbol{U}_{(i)}^T\boldsymbol{U}_{(i)}\right]$, with a complexity of $\mathcal{O}(\ell^3)$ operations. Fortunately $\boldsymbol{C}_{(i)}$ can be written as a rank one modification of $\boldsymbol{C}$,

$$\boldsymbol{C}_{(i)} = \left[\boldsymbol{R}_{(i)} + \boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{u}_i\boldsymbol{u}_i^T\right] = \left[\boldsymbol{C} - \boldsymbol{u}_i\boldsymbol{u}_i^T\right], \qquad (5)$$

where $\boldsymbol{u}_i$ is the i^{th} row of $\boldsymbol{U}$. This allows $\boldsymbol{C}_{(i)}^{-1}$ to be found in only $\mathcal{O}(\ell^2)$ operations [15], given that $\boldsymbol{C}^{-1}$ is already known, via the following matrix inversion formula : Given an invertible matrix $\boldsymbol{A}$ and column vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, then assuming $\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u} \neq -1$, we have that

$$\left(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T\right)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}}.$$

The computational complexity of the matrix inversion at each step is thus reduced from $\mathcal{O}(\ell^3)$ to $\mathcal{O}(\ell^2)$. The computational complexity of the leave-one-out cross-validation process is then only $\mathcal{O}(\ell^3)$ operations, which is the same as that of the basic training algorithm for the kernel Fisher discriminant classifier. However, a further refinement is possible, it can be shown [16] that the leave-one-out error $E_{loo} = E_{loo}(\{\boldsymbol{r}_{(i)}\}_{i=1,\ell}, \boldsymbol{y})$, can be computed analytically in closed form using

$$\left\{\boldsymbol{r}_{(i)}\right\}_i = \frac{1}{1 - h_{ii}} r_i.$$

where $\left\{\boldsymbol{r}_{(i)}\right\}_i = y_i - \boldsymbol{w}_{(i)} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) - b_{(i)}$ is the residual error for the i^{th} training pattern during the i^{th} iteration of the leave-one-out cross-validation procedure, $r_i = y_i - \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) - b$ is the residual error for the i^{th} training pattern for a kernel Fisher discriminant classifier trained on the entire dataset, and $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{C}^{-1}\boldsymbol{U}^T$ is the *hat* matrix of which h_{ii} is the i^{th} element of the leading diagonal [14]. In this case, $\boldsymbol{C}$ is diagonal and can be inverted in linear time, and therefore

$$h_{ii} = \sum_{j=1}^{\ell} u_{ij}^2 c_{jj}^{-1} = \sum_{j=1}^{\ell} \frac{u_{ij}^2}{(\lambda_j + \mu_j)}.$$

The leave-one-out error rate can thus be evaluated in closed form without explicit inversion of $\boldsymbol{C}_{(i)} \ \forall i \in \{1, 2, \ldots, \ell\}$, with a computational complexity of only $\mathcal{O}(\ell^2)$ operations. To find the optimal regularisation parameters we will assume, as is normally the case, a single regularisation parameter μ, the optimal value, minimising the leave-one-out error, is then found using a simple line search.

III. Identification of Putative Regulatory Motifs

Three different feature extraction methods were used to extract motifs corresponding to putative transcription factor binding sites (regulatory motifs). The first approach simply took sequences from the PLACE database [17] representing experimentally determined plant cis-acting regulatory elements. These are sequences that are known to influence regulation in a variety of plants under a variety of stimuli (i.e. not necessarily implicated in glucose response). The second approach generated a small number of motifs using a partially automated "heuristic" approach, involving some intervention from the investigator. Lastly, the spectrum and mismatch kernels were used to implicitly create classifiers in a kernel-induced feature space comprised of all possible k-mers with up to m mismatched symbols. These approaches employ varying degrees of expert knowledge, from PLACE (high) to mismatch kernel (low), varying numbers of features, from mismatch kernel (high) to heuristic (low), and provide varying degrees of interpretability, from PLACE (high) to mismatch kernel (low).

A. Features Extracted from the PLACE Database

PLACE[1] is a database of motifs representing plant cis-acting regulatory DNA elements that have been obtained through experiments described in previously published reports on genes (principally) in vascular plants [17]. These sequences represent regulatory elements from a variety of plants, controlling regulatory response to a variety of stimuli, some of which relate to specific parts of the plant. No attempt was made to account for genetic divergence between *Arabidopsis thaliana* and the other plants targeted by entries in the PLACE database, which include rice (*Oryza sativa*), maize (*Zea mays*), tomato (*Lycopersicon esculentum*) and wheat (*Triticum aestivum*). Note also that many of the elements described in the PLACE database have not previously been implicated in glucose-response in any plant. A matrix was constructed, each column of which gives the number of occurrences of the sequence representing a PLACE element, or its complement, in the promoter of every gene co-regulated in response to glucose. Of the 381 PLACE elements, only 253 were found to occur in the promoters of the genes included in this study.

B. Features Extracted via Heuristic Search

A heuristic search, guided by the investigator, was used to test whether the use of a highly compact feature set substantially improved or degraded performance. The search began with the set of 4^5 distinct $5-$mers drawn from the alphabet $\{\mathtt{A,C,G,T}\}$. As we count a each 5-mer and its complement as being the same feature (as we do not distinguish between conformations of the double-helix that differ only in their orientation with respect to the transcribed region), and so we discard any sequence that is lexically greater than its complement. The remaining sequences were then scored according to a commonly used correlation coefficient (e.g. [18]),

$$f(x_j) = \frac{\mu^+ - \mu^-}{\sigma^+ + \sigma^-} \qquad (6)$$

where x_j is the j^{th} motif, μ^+ and μ^- represent the mean number of occurrences of the j^{th} in the promoters of genes in the positive (e.g. up-regulated) and negative (e.g. down-regulated) classes, and and σ^+ and σ^- are the corresponding standard deviations. This formula is related to the criterion used in Fisher's linear discriminant analysis [7, 8] and a high positive or negative value indicates a highly discriminant feature. A three-way comparison was performed, investigating up-versus-down, up-versus-unregulated and down-verses-unregulated sets of genes. The features with high scores on up-versus-down, up-versus-unregulated *and* down-versus-unregulated were rejected as being equivocal. Next any motif with low coverage (below 30%) in either the up- or down-regulated sets were discarded, before selecting the motifs achieving correlation scores in the top 10% in the comparison of up-versus-down and up-versus-unregulated genes. At this point, five 5-mers associated with enhanced glucose response (AAACC, AACCC, ACCCT, CCCTA and CTACT) and eighteen 5-mers associated with glucose suppression (AAGAT, AATAT, ACGTG, AGATA, ATCAT, ATCCA, ATTAT, CACAT, CCACT, CTATC, GATAA, GATAT, TAAAG, TACGT, TATCC, TATCT, TATTA and TATAC) had been identified).

TABLE I
IUPAC WILDCARD SYMBOLS USED IN ADDITION TO A, C, G AND T.

Symbol	Bases	Symbol	Bases
B	C, G or T	D	A, G or T
H	A, C or T	K	G or T
M	A or C	N	A, C, G or T
R	A or G	S	C or G
V	A, C or G	W	A or T
Y	C or T		

Motifs were then grouped, such that any two 5-mers that share at least four consecutive bases were combined to form a 4-mer, e.g. AAACC and AACCC were combined to form AACC. Many transcription factor binding sites are assumed to be composed of a "core" sequence providing the bulk of its specificity, surrounded on both sides by less specific "flanking" sequences. The remaining features were then augmented by 3 base pairs of up-stream and down-stream flanking sequences. These were formed by an analysis of the regions immediate up- and down-stream of matches between the promoters of the co-regulated genes and the "core" sequence. IUPAC wildcard symbols (Table I) were added to the core motifs to accommodate any over-represented nucleotide in any of the six flanking positions. This resulted in the final set of 11 motifs : AAACCCTAA and CTACT associated with up-regulated genes and AAGATAW, YACGTG, YTATCYA, TATTAT, AATAT, ATCAT, CACAT, CCACT, TAAAG associated with down-regulated genes. It is interesting to note that two of these motifs are also found in the PLACE database, namely AAACCCTAA, known as the "TELOBOX" element and YACGTG forming a substantial part of the "ABREATCONSENSUS" element.

C. Features "Extracted" from Sequence Kernels

Kernel learning methods have been found to be particularly well suited to many problems arising in computational biology [19] as it is relatively straight-forward to construct kernel functions operating directly on structured data, for instance variable length sequences of symbols drawn from a fixed

[1] Available from http://www.dna.affrc.go.jp/htdos/PLACE.

alphabet, such as DNA sequence data. In this study, we use two such kernel functions, the k-spectrum kernel the k-spectrum kernel [20] and the closely related $(k$-$m)$-mismatch kernel [21, 22]. The feature space of the k-spectrum kernel records the number of occurrences of all possible substrings of length k from an alphabet $\mathcal{A}$ found in string, $\boldsymbol{x}$, i.e.

$$\boldsymbol{\Phi}_k(\boldsymbol{x}) = (\phi_a(\boldsymbol{x}))_{a \in \mathcal{A}^k}$$

where $\phi_a(\boldsymbol{x})$ gives the number of times the substring a occurs in $\boldsymbol{x}$. The k-spectrum kernel, which computes the inner product between vectors in the space of all possible k-mers,

$$\mathcal{K}_k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\Phi}_k(\boldsymbol{x}) \cdot \boldsymbol{\Phi}_k(\boldsymbol{x}')$$

then measures the similarity of a pair of strings in terms of the "density" of shared substrings of length k. The (k, m)-mismatch kernel,

$$\mathcal{K}_{(k,m)}(\boldsymbol{x}, \boldsymbol{x}'),$$

extends the k-spectrum kernel, by allowing up to m mismatches in the determination of the set of shared substrings [20]. The feature space is then defined as follows : Let α, β represent k-mers in $\mathcal{A}$, then

$$\boldsymbol{\Phi}_{(k,m)}(\alpha) = (\phi_\beta(\boldsymbol{x}))_{\beta \in \mathcal{A}^k}.$$

where $\phi_\beta(\boldsymbol{\alpha})$ is 1 if the k-mers α and β differ in at most m locations and 0 otherwise. The feature vector for the entire string $\boldsymbol{x}$ is then found by summing over all substrings of length k occurring in $\boldsymbol{x}$,

$$\boldsymbol{\Phi}_{(k,m)}(\boldsymbol{x}) = \sum_{\alpha \in \boldsymbol{x}} (\boldsymbol{\Phi}_{(k,m)}(\alpha)).$$

The k-spectrum and $(k$-$m)$-mismatch kernels allow us to implicitly construct classifiers in the space of all possible substrings of length k, possibly allowing up to m mismatches to account for variation in transcription factor binding sites in genes with different evolutionary paths. Importantly, these kernels place no limitations on the initial set of putative regulatory motifs, but also incorporate very little expert knowledge.

A third kernel used in this study is the inhomogeneous polynomial kernel

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x} \cdot \boldsymbol{x}' + 1)^p\,,$$

which induces a feature space comprised of products of all combinations of p or less of the input variables. The use of this kernel allows us to implicitly include features representing *combinations* of regulatory motifs, without incurring the computational expense in evaluating these product features explicitly.

IV. Results

A database was assembled comprising approximately 1000 b.p. of 5' flanking sequences of 1051 genes with greater than 2.5 fold increase in response to glucose at 2, 4 and 6 hrs (the "Up" set), 793 promoters of genes with reduced expression in response to glucose at 2, 4 and 6 hrs (the "Down" set) and 964 un-regulated genes (the "Neutral" set) for this work. The experimental results presented in this section are concerned with distinguishing between up- and down-regulated sets of genes. Table II shows the leave-one-out cross-validation error rates for ORKFD classifiers, based on a polynomial kernel, for feature sets derived from PLACE elements and extracted via the heuristic search procedure. Two features are immediately apparent: Firstly the feature set derived from the set of PLACE elements out-performed the feature set extracted via heuristic search, suggesting that the heuristic search procedure did not extract all of the useful discriminatory motifs from the promoters. Secondly, in both cases, the performance deteriorated as the order, p, of the polynomial kernel increased. This suggests that the discriminatory a combination of motifs is not greater than the sum of their parts, providing a useful insight into the co-ordination of gene regulation. Experiments using both the PLACE and heuristic feature sets demonstrated the "TELOBOX" motif to be discriminative in distinguishing up- from down-regulated genes. This is interesting as this regulatory element has not previously been implicated in sugar-regulation. We have since verified this result experimentally *in vivo*, demonstrating that our approach can be used to extract novel biological knowledge from microarray data.

TABLE II
Leave-one-out cross-validation error rates obtained for ORKFD classifiers, based on a polynomial kernel, using features derived from the PLACE database and features extracted via heuristic search.

p	PLACE	Heuristic
1	**0.33**	**0.34**
2	0.47	0.36
3	0.48	0.39
4	0.49	0.43
5	0.49	0.47
6	0.48	0.48
7	0.47	0.47
8	0.47	0.49
9	0.48	0.48
10	0.46	0.48

Table III shows the leave-one-out cross-validation error for ORKFD classifiers based on the $(k$-$m)$-mismatch kernel, for various values of k and m. The best classifier is obtained for $(k = 4, m = 1)$ suggesting that over-fitting becomes more difficult to prevent for very precise feature sets (i.e. as k becomes large). However, the best classifier out-performs classifiers based on PLACE and heuristic search feature sets, suggesting that there are regulatory elements not well represented by the latter. Note also that for small k, longer regulatory motifs may be represented by $(k$-$m)$-mismatch features may be encoded by a pattern of activation over a number of length k features.

V. Conclusion

In this paper, we have applied a new kernel learning method, namely the optimally regularised kernel Fisher discriminant

TABLE III
Leave-one-out cross-validation error (LOOCVE) rates obtained for ORKFD classifiers, based on the $(k\text{-}m)$-mismatch kernel.

k	m	LOOCVE
4	1	**0.32**
5	1	0.34
5	2	0.34
6	1	0.38
6	2	0.38
6	3	0.38
7	1	0.37
7	2	0.37
7	3	0.37
8	2	0.38

(ORKFD) classifier, for promoter-based gene classification. The ORKFD provides an efficient means to set the regularisation parameter so as to minimise the leave-one-out cross-validation error. This makes the ORKFD an attractive tool for applications in computational biology as it not only avoids over-fitting, but also greatly simplifies the model selection procedure, where only the values of a small number of typically discrete kernel parameters remain to be found. The analysis of glucose response in *Arabidopsis thaliana* has revealed a novel role for the "TELOBOX" regulatory element, that had not previously been implicated in glucose response - a finding that has since been verified *in vivo*. The study has also suggested that, although regulatory elements act in combination to co-ordinate gene expression, the discriminatory power of a combination of motifs is not greater than the sum of the individual elements, so a linear learning method should suffice.

VI. Acknowledgements

This work was supported by a grant from the Biotechnology and Biological Sciences Research council (BBSRC), grant number 83/EGM16128.

References

[1] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines (and other kernel-based learning methods)*. Cambridge University Press, Cambridge, U.K., 2000.

[2] B. Schölkopf and A. J. Smola. *Learning with kernels - support vector machines, regularization, optimization and beyond*. MIT Press, Cambridge, MA, 2002.

[3] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[4] A. Zien. A primer on molecular biology. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel methods in computational biology*, chapter 1, pages 3–34. MIT Press, 2004.

[5] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Essential cell biology : An introduction to the molecular biology of the cell*. Garland Science, 1997.

[6] K. Saadi, N. L. C. Talbot, and G. C. Cawley. Optimally regularised kernel Fisher discriminant analysis. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR-2004)*, volume 2, pages 427–430, Cambridge, United Kingdom, August 23–26 2004.

[7] C Bishop. *Neural Networks for Pattern Recognition*. Oxford university Press, Oxford, UK, 1995.

[8] A. Webb. *Statistical pattern recognition*. Wiley, second edition, 2002.

[9] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing*, volume IX, pages 41–48. IEEE Press, New York, 1999.

[10] J. Xu, X. Zhang, and Y. Li. Kernel MSE algorithm: A unified framework for KFD, LS-SVM and KRR. In *Proc. IJCNN*, pages 1486–1491, Washington, DC, July 2001.

[11] A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed problems*. John Wiley, New York, 1977.

[12] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilema. *Neural Computation*, 4(1):1–58, 1992.

[13] A. Luntz and V. Brailovsky. On estimation of characters obtained in statistical procedure of recognition (in russian). *Techicheskaya Kibernetica*, 3, 1969.

[14] S. Weisberg. *Applied linear regression*. John Wiley and Sons, New York, second edition, 1985.

[15] M. Woodbury. Inverting modified matrices. Memorandum report 42, Princeton University, Princeton, U.S.A., 1950.

[16] G. C. Cawley and N. L. C. Talbot. Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers. *Pattern Recognition*, 36(11):2585–2592, November 2003.

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

[18] T. Golub and *etal.* Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[19] B. Schölkopf, K. Tsuda, and J.-P. Vert. *Kernel methods in computational biology*. MIT Press, Cambridge, MA, 2004.

[20] C. Leslie, E. Eskin, and W. Stafford Noble. The spectrum kernel : A string kernel for SVM protein classification. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 564–575, January 2–7 2002.

[21] C. Leslie, E. Eskin, J. Weston, and W. Stafford Noble. Mismatch string kernels for SVM protein classification. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1417–1424. MIT Press, Cambridge, MA, 2003.

[22] C. Leslie, E. Eskin, A. Cohen, J. Weston, and W. Stafford Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–476, 2004.