# Classification of co-expressed genes from DNA regulatory regions

Giulio Pavesi [a] and Giorgio Valentini [b,*]

[a] *Dipartimento di Scienze Biomolecolari e Biotecnologie, Università degli Studi di Milano, Via Celoria 20, Milano, Italy.*

[b] *Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Via Comelico 39, Milano, Italy.*

**Abstract**

The analysis of non–coding DNA regulatory regions is one of the most challenging open problems in computational biology. In this paper we investigate whether we can predict functional information about genes by using information extracted from their sequences together with expression data. We formalize this problem as a classification problem, and we apply Support Vector Machines (SVMs) with non linear kernels to predict classes of co-expressed genes obtained from clustering procedures. SVMs are trained using information about selected motifs extracted from DNA regulatory regions through combinatorial and statistical methods. In our experiments, we show that functional classes of genes can be predicted from biological sequence data in *S. cerevisiae*, achieving results competitive with those recently presented in the literature.

*Key words:* Gene classification, motif extraction and selection, Gene expression and bio-sequence data integration, Combinatorial and machine learning methods integration

## 1 Introduction

Genome sequencing projects have produced the full DNA sequence of human and a number of other organisms, opening new avenues for research in biology and medicine. Fundamental units of the information encoded in a genome

* Corresponding author.
*Email addresses:* giulio.pavesi@unimi.it (Giulio Pavesi), valentini@dsi.unimi.it (Giorgio Valentini).

are *genes*. Roughly speaking, a gene can be described as a specific region of double–stranded DNA which is transcribed into a single–stranded RNA sequence, which is in turn translated into a protein [1] (see Fig. 1). While the annotation of the whole gene repertoire of the different genomes available is still ongoing, a striking feature that has emerged is the fact that genes are not all simultaneously activated and translated into a protein (*expressed*); rather, specific subsets of genes are active at any given time in a given cell, according for example to the type of cell itself (the genes that are active in a neural cell are different from those active, say, in a muscle cell - thus leading to cell differentiation) or to external stimuli. As a matter of fact, genetic diseases are often caused by alterations occurring not within the genes themselves and the protein they encode, but in the apparatus governing their activation, thus leading to anomalous expression levels. Thus, one of the main challenges in modern biology in general, and in the analysis of genomic data in particular, is to understand the complex mechanisms that regulate the expression of the genes of a given organism.

An important step toward this direction is the identification of the motifs (short oligonucleotide sequences) responsible for the binding of transcription factors (TF) that regulate the expression of genes. A common computational approach to this problem consists in looking for overrepresented sequence motifs in DNA regulatory regions of co-expressed genes [2]. Unfortunately these statistical methods are in some cases unsuccessful, because the degree of conservation in binding sites for the same TF is often very low, and because co-expression is in general not synonymous with co-regulation [3].

To address these problems, in this paper we apply a machine learning approach to classify sets of co-expressed genes using information extracted from their regulatory regions. Our goal is to investigate whether functional information about genes can be inferred from the sequences of the corresponding DNA regulatory regions, and the motifs extracted from them. Indeed, if classes of co-regulated genes can be predicted from their regulatory motifs, we can indirectly gain information about their biological significance. To this end, we propose a computational approach that integrate heterogeneous bio-molecular data with a "heterogeneous" learning system. More precisely, on the one hand we adopt a "data fusion" approach integrating numerical gene expression data of co-regulated genes with sequences extracted from their corresponding DNA regulatory regions. On the other hand, from an algorithmic viewpoint, we combine machine learning methods to classify classes of co-expressed genes with combinatorial algorithms to extract motifs from sequence data, and with statistical methods to select subsets of significant motifs.

Within this proposed integrated learning system, Support Vector Machines (SVMs) play a central role in answering to the main question arisen from this work: given that specific regulatory regions are responsible for the regulation

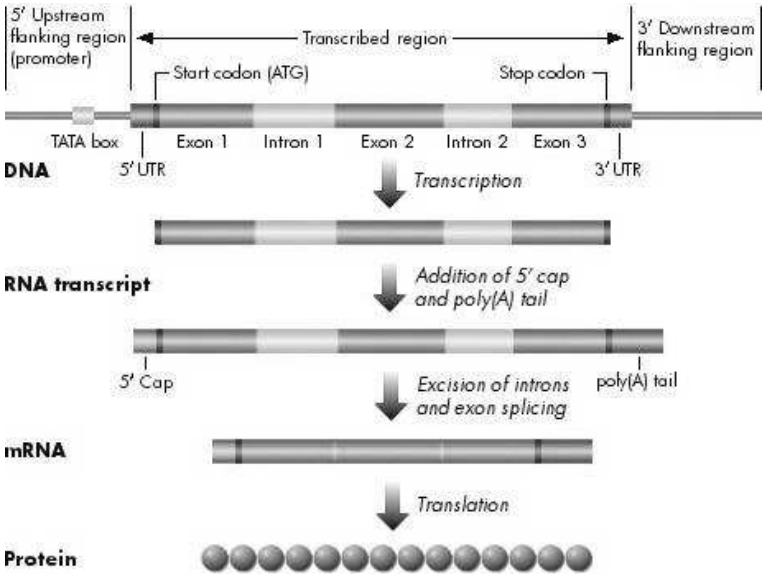



Fig. 1. Gene expression: from DNA to RNA to protein.

of gene expression, to which degree gene expression can be predicted from those local regulatory DNA sequences?

In the next section we discuss more in detail the biological problem of gene expression regulation and some of the main computational approaches to this problem. Then in Sect. 3 we introduce our method. In Sect. 4 we briefly discuss commonalities and differences with other methods related to our work. Finally, we present experimental results we obtained from the analysis of promoter sequences in yeast, showing that they compare favourably with recent literature presenting similar techniques.

## 2 Computational approaches to the analysis of gene expression regulation.

The expression of a gene starts when the corresponding region in the double–stranded DNA sequence is *transcribed* into a single stranded RNA sequence. Transcription is initiated when one or more dedicated molecules called *transcription factors* (TFs) (that are proteins in turn encoded by some genes in the genome) bind to the DNA region adjacent to the gene (region called *promoter* of the gene), causing the double–strand to open and thus allowing the transcription of the gene. In some cases, the binding of a TF has the opposite effect, blocking (or, silencing) transcription. Each TF recognizes a set of specific targets along the sequence, that is, short nucleotide fragments it can bind to, called *binding sites* (TFBSs). Binding sites thus function as regulatory signals in the genome, that can be seen as buttons, that can be pushed by the different TFs recognizing them. When all the buttons are pushed, by

the right TFs at the right time, transcription starts.

Since the experimental in vivo and in vitro characterization and identification of the binding sites for a given TF is a long and painstaking work, the huge amount of genomic data now available to researchers provides a invaluable source of information for shedding further light on this process. If we identify in the promoter of a gene known TFBSs, we may better understand by what, and when a gene is activated. Unfortunately, also the computational description and discovery of the binding sites of a given TF is far from being an easy task. The main difficulty lies in the very fact that each TF does not recognize a single binding site, but a set of them, that, although similar, differ in their nucleotide composition. This set of similar DNA words recognized by a TF is usually referred to as *signal* or *motif*. A typical TFBS motif is usually 6–15 nucleotides long.

On the other hand, recent lab techniques like microarray chips have allowed researchers to gather data about the simultaneous expression level of several genes under different conditions [4]. Briefly, the expression level of the genes is measured according to how much transcripts they produce. For each gene, two different experimental conditions are usually compared, measuring the difference in the transcript levels produced. Then, clustering algorithms are usually applied to identify groups of co-expressed genes, that is, genes whose expression varies in a similar fashion in two or more different experiments [5]. Genes belonging to the same cluster are thus very likely to be co-regulated, that is, activated by the same set of TFs [6].

Thus, a very common computational approach to the problem is, given a set of co-expressed or co-regulated genes, to look for conserved motifs in their regulatory regions, usually the promoters. Motifs represent the binding specificity of the TF(s). In other words, we expect to find in each (or most) of the regions short DNA fragments (called oligonucleotides or shortly oligos), similar enough to each other to be considered as instances of binding sites for the same TF. Computational methods for this task are thus based on two steps: first, one or more than one group of similar oligos are detected, and, second, their presence/conservation is evaluated from a statistical point of view, that is, how likely would be the same group to be detected with the same degree of conservation in a set of random sequences, i.e., from genes not co-regulated [2].

As different tests have shown [3], the problem formalized in this way is very hard, given the often low degree of conservation in sites for the same TF. Furthermore, what researchers often have at their disposal is a set of co-expressed genes (e.g. obtained from a microarray experiment). Although subtle, the difference is important: co-regulation means having the same TF(s) regulating all the genes, while co-expression implies that several different TFs can be involved in the regulation, each one on a subset of the genes, and thus in turn

conserved motifs do not appear in every input sequence and their statistical significance often is too low to discriminate them from random similarities.

All in all, the result is that available methods usually output a long list of motifs, containing inevitably many false positives, leaving researchers with several candidates to test. Furthermore, while in simple eukaryotes like yeast the promoter sequence very often contains all the TFBSs that regulate the transcription of a gene, in higher organisms, including naturally human, promoter sequences alone often are not enough to fully explain the regulation of a gene, that can be influenced by TFBSs located within distal elements like enhancers or silencers that can be situated even at several thousands of base pairs from the gene itself; and motif discovery algorithms do not give any information whether motifs found are sufficient to explain the co-regulation (or co-expression) of the genes investigated. Another important issue is, given the motifs extracted from the sequences, to determine whether additional genes, for which for example expression data are not available, can be predicted to have the same expression profile. In this case, the traditional approach is locate in their promoter or other regulatory regions one or more motifs, and to give a statistical estimate of the significance of finding them: however, it is very hard to have a feasible prediction of the expression of a gene from motifs occurrences alone.

Researchers, however, often have at their disposal more information than a single set or cluster of co-expressed genes. From the analysis of a microarray experiment, for example, it is common to have different clusters of genes, expressed in different ways: thus, additional help in the analysis could be gained by the comparison of clusters of genes with different expression profile, rather than the analysis of a single one. Traditional approaches can be modified to take this into account, merging numerical data concerning gene expression with sequence analysis, by looking for motifs over–represented in a set and at the same time under–represented in other set (see for example [7]), but the issues just outlined remain open.

## 3 Combining combinatorial, statistical and machine learning methods for motif-based prediction of genes

The basic idea behind our approach is that, given a set of sequences (e.g promoters) from a set of co-expressed genes containing the TFBSs responsible for the regulation, we should be able to train a classifier that based on sequence information alone would discriminate genes with different expression pattern or simply picked at random from the same organism. Since regulation is based on common TFBSs (motifs), then one possible approach is first to extract motifs from the sequences of the training set, then to train the classifier on

the motifs extracted. This, in turn, would allow us to define which minimal motif set is sufficient for obtaining successful predictions, in other words, to explain the co-expression (or co-regulation) of the genes. Also, failure in the training or in the testing of the classifier could imply that the sequences analyzed are not enough; in other words, that other elements not included in the input sequences are involved in the regulation of the genes and responsible for their co-expression. For example, failure to predict tissue–specific expression for human genes based on promoter sequence alone might imply that other elements (like enhancers) could be involved.

Given a set of co-expressed genes (the positive set) and another one with different expression pattern or composed of genes picked at random from the same organism (the negative set), the approach we propose merges a combinatorial approach aimed at the extraction of motifs from the input sequences with statistical methods for the identification of the most relevant motifs, and with machine learning techniques, integrating sequence data (the promoter sequences from the genes) with numerical expression data, used to identify clusters of co-expressed genes. Our algorithm can be thus split into two major points:

(1) Motif extraction and selection: motifs are extracted and scored from sequences (e.g. promoters) of genes in both the positive and the negative set. Then using the previously computed scores, the most significant motifs are selected.
(2) Gene classification: a classifier is trained, on the basis of the motifs selected from the two sets at the previous step.

The different steps of our method are summarized in Figure 2.

The trained classifier should be able to predict whether the expression of one or more additional genes can be associated with either cluster. Furthermore, our idea is to find motifs that are sufficient to discriminate one cluster from the other: in this way, a minimal set of motifs can be identified. Thus, as stated before, information on which motifs are actually responsible for the expression of the genes could be gathered.

In the next subsections we detail our motif extraction, selection and gene classification procedures.

### *3.1 Motifs extraction and selection*

For the first step, extracting motifs from the input sequences, several different methods and approaches have been proposed [2]. Roughly speaking, motif discovery algorithms can be split into two broad categories: exhaustive or heuris-

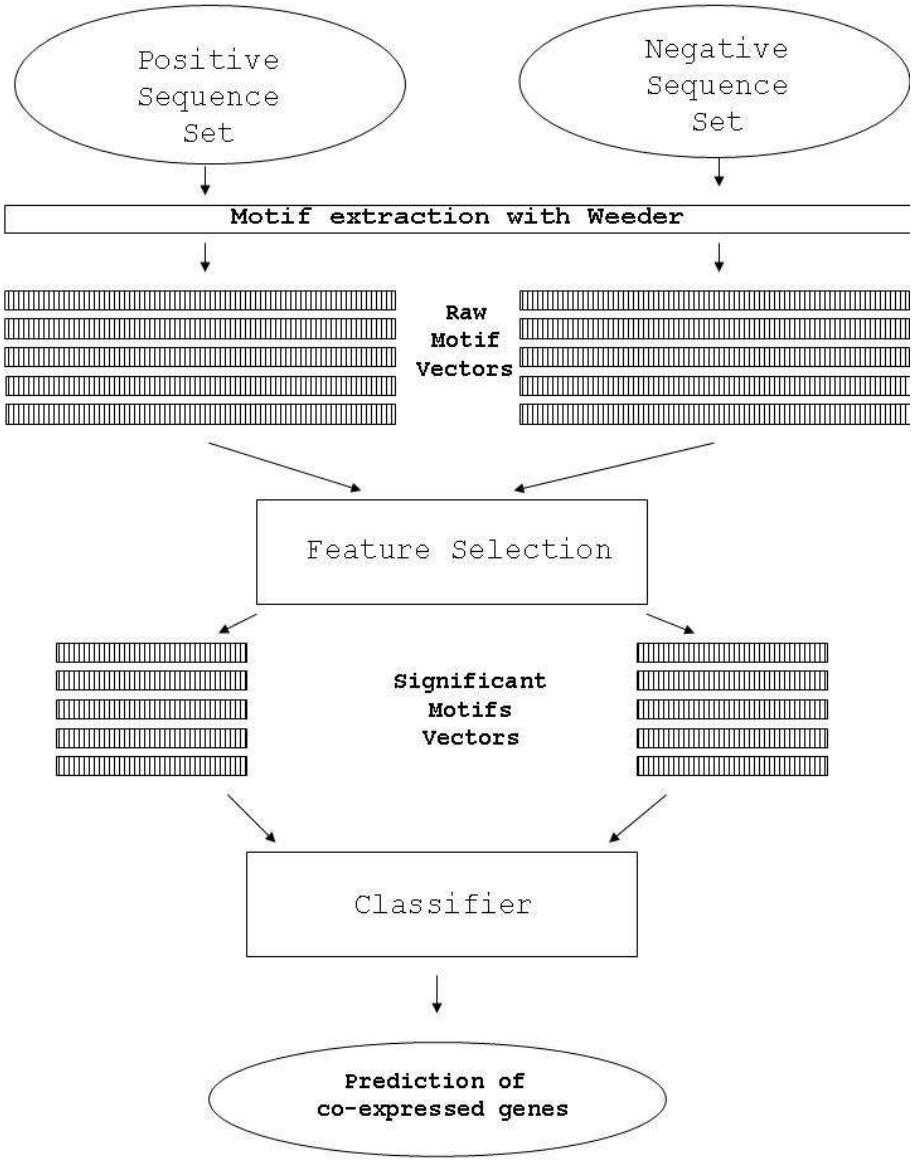



Fig. 2. The structure of the algorithm. Each sequence is converted into a motif vector representing the score contribution of the sequence to each of the $4^m$ possible motifs of length $m$. Feature selection techniques are applied to the vectors, in order to select the most significant motifs for either cluster. A classifier is then trained on the resulting vectors.

tic methods. In the former, the algorithms evaluate the statistical significance of all possible motifs, and output a ranked list. We deemed this approach very suitable for our task, since it spares from the need of pre–selecting a subset of motifs to use in the classification. In other words, we wanted to consider every possible motif, so to be able to identify subtle motifs appearing for example in a limited subset of genes but nevertheless able to characterize one cluster of genes with respect to the other.

Thus, for this task, we employed the Weeder algorithm [8], that has also the merit of achieving better performances than most of the other methods introduced for the same task [3,9]. Weeder represents motifs with a consensus, describing the most frequent nucleotide appearing in each position of the different motif instances. All oligos differing from the consensus in a limited number of substitutions can be thus a priori considered as valid motif instances.

The basic idea of the algorithm is that in a set of promoters from co-regulated genes, one should detect one or more oligos (motifs) appearing (approximately) in the sequences a number of times significantly higher than the one that would be obtained from a set of sequences picked at random from the same organism. For this task, for each oligo of suitable size the algorithm uses pre–computed frequencies obtained from the analysis of all the promoter sequences of different species. Thus, the number of occurrences of each oligo is compared to the expected value, as follows.

Chosen a length $m$ for the motifs and a set of $k$ input sequences, all the $4^m$ possible oligos are scored, according to their number of occurrences in the sequences and their conservation. The basic Weeder score for motif $M$ is defined as

$$W(M) = \sum_{i=1}^{k} I(i, e) \log \frac{Occ(i, e_i)}{Exp(i, e_i)}$$

where $e$ is the maximum number of substitutions allowed in the occurrences of the motif; $I(i, e)$ equals 0 if $M$ does not appear in sequence $i$ with at most $e$ substitutions, 1 otherwise; $e_i$ is the minimum number of substitutions with which $M$ appears in sequence $i$; $Occ(i, e_i)$ is the number of times $M$ appears in sequence $i$ with $e_i$ substitutions; and finally $Exp(i, e_i)$ is the expected value for $Occ(i, e_i)$ that can be computed by performing an oligo analysis of all the sequences of the same type (e.g. promoters) from the same species of the input organism [8].

Notice that the score can be split into $k$ terms, one for each sequence analyzed. Moreover, given motif length $|M|$, this score is computed for all the possible $4^{|M|}$ motifs. The result is that we can define for each input sequence a vector of $4^{|M|}$ elements, denoting the score contribution given by the sequence on the overall score of each motif. In the experiments we present in this article, we ran Weeder on motifs of length 6 and 8 (sixmers and eightmers), allowing at most 1 substitution in their occurrences. The result obtained from the first phase is thus a $4^6$ or $4^8$ elements vector for each input sequence.

It is worth noting that the Weeder algorithm, using suffix trees, has a time complexity linear in the overall length of the analyzed sequences, thus allowing an efficient computation of the motif scores [8].

The next step consists in identifying subsets of motifs characteristic of specific sets of co-expressed genes (expression patterns), when compared to each other.

This problem can be formalized in different ways: we chose a statistical approach, since it permits the computation of significance levels to be associated with the motif scores. Therefore, we applied a statistical test of hypothesis to select:

(1) Subsets of the most significant motifs related to a specific set $\mathcal{C}_a$ of co-expressed genes versus another $\mathcal{C}_b$, $b \neq a$.
(2) Subsets of the most significant motifs related to a specific set $\mathcal{C}_a$ versus sets of genes randomly drawn from the same organism.

For each $n$-dimensional motif vector of Weeder scores $\mathbf{x}_j = (x_{j1}, \ldots, x_{jn})$, $1 \leq j \leq m$, where $n$ is the number of genes, $m$ is the number of motifs, and $\mathbf{y} \in \{+, -\}^n$ is the vector of the labels of the sets of genes, we computed the corresponding Welch t-statistic $T_j$ to take into account possible differences in the variance of motif-scores between the two sets of genes:

$$T_j = \frac{\bar{x}_j^+ - \bar{x}_j^-}{\sqrt{\frac{s_j^+}{n^+} + \frac{s_j^-}{n^-}}} \tag{1}$$

where $\bar{x}_j^+$ and $\bar{x}_j^-$ are the sample means of the the positive and negative sets of genes for the $j^{th}$ motif and $s_j^+$ and $s_j^-$ the corresponding sample variances. Using Eq. 1 we applied the two–sample Welch t-test to verify the null hypothesis $\mathcal{H}_j$ of no difference between the means of motif scores of the two given positive and negative sets of $n$ genes at a given significance level $\alpha$. In this way we can obtain a subset $J \subset \{1, \ldots, m\}$ of motifs selected at $\alpha$-significance level, that can be used to classify the classes of co-regulated genes:

$$J = \{j || T_j | > t_{\alpha/2, n-1}\} \tag{2}$$

where $t_{\alpha,n}$ is defined according to the integral of the *Student* $t$-distribution $f(t)$ with $n$ degrees of freedom:

$$\int_{t_{\alpha,n}}^{\infty} f(t)dt = \alpha \tag{3}$$

Note that our approach implicitly assumes that there are no correlations between motifs. This is clearly an oversimplification, since motifs for co-operating or competitive factors are often found to co-occur in regulatory sequences. Anyway, considering that in general the interactions are quite limited (considering the overall number of motifs involved), and the signals (motifs) are

quite noisy, with a certain approximation the independence assumption can be considered acceptable.

### *3.2 Genes classification*

Using the subset of motifs selected according to the procedures described in the previous section, we obtain the examples $\mathbf{z}_i, 1 \leq i \leq n$ (one example for each gene) with only the Weeder scores relative to the subset of selected motifs (eq. 2):

$$\mathbf{z}_i = \{x_{ji} || T_j | > t_{\alpha/2,n-1}\} \tag{4}$$

where $x_{ji}$ is the weeder score of the $j^{th}$ motif of the $i^{th}$ gene. These data are then used to classify classes of co-regulated genes.

For this difficult task we chose soft margin Support Vector Machines (SVMs). The SVM algorithm minimizes both the empirical risk and the margin between the convex hulls of the two classes, thus assuring at the same time an accurate learning of the training data and generalization capabilities [10]. In our case the classification problem is highly non linear and then we need to apply kernels to perform a linear separation in the corresponding inner product feature space. In the dual formulation of the constrained quadratic program underlying SVMs, we need to solve the following optimization problem:

$$\begin{aligned} \text{Maximize} \quad & \Phi(\alpha) = \textstyle\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(\mathbf{z}_i \mathbf{z}_j) \\ \text{subject to} \quad & \textstyle\sum_{i=1}^{n} y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad 1 \leq i \leq n \end{aligned}$$

where $n$ is the number of examples, $K(\mathbf{z}, \mathbf{z}')$ is a symmetric function satisfying *Mercer's conditions* [11], $C$ is the regularization parameter controlling the trade-off between the accuracy on the training set and the width of the margin, while $\alpha_i$ are the Lagrange multipliers raised from the solution of the primal constrained optimization problem.

The discriminant function obtained from the solution $\alpha^*$ of this quadratic optimization problem is:

$$f(\mathbf{z}, \alpha^*, b^*) = \sum_{i=1}^{n} y_i \alpha_i^* K(\mathbf{z}_i, \mathbf{z}) + b^*$$

The symmetric function $K(\cdot, \cdot)$ must be chosen among the kernels of Reproducing Kernel Hilbert Spaces; we chose:

- Gaussian kernels: $K(\mathbf{u}, \mathbf{v}) = \exp(-\|\mathbf{u} - \mathbf{v}\|^2/\sigma^2)$
- Sigmoid kernels: $K(\mathbf{u}, \mathbf{v}) = \tanh(c_1(\mathbf{u} \cdot \mathbf{v}) + c_2)$

where $\sigma$, $c_1$ and $c_2$ are tunable kernel parameters.

## 4 Related work

Motif extraction and classification of sequences based on the motifs is an approach that has already been introduced. In the seminal work of Beer and Tavazoie [12], for example, motifs representing known instances of TFBSs coupled with novel motifs extracted from promoter sequences are used to predict the expression profile of yeast genes by training a Bayesian network on 49 gene clusters obtained from cell-cycle and environmental stress microarray data. The overall classification accuracy was around 73%, showing that indeed it was possible to predict gene expression from sequence information alone, provided that the latter contains all the information needed. Anyway it is worth noting that the authors used a priori known regulatory motifs for their predictions, while in our approach we did not exploit any a priori biological knowledge about known motifs.

The same objection can be applied to [14], that employs models of known sites for TFs responsible for tissue–specific expression to predict tissue–specific genes in human and mouse, with an accuracy around 60–65% in most of the tissues considered.

In [13], instead, motifs are extracted by using sequence and position conservation in order to characterize signals in core promoters and introns for promoter and splicing prediction by training a Support Vector Machine.

Recently, an approach similar to ours, implemented with a Relevance Vector Machine (RVM) coupled with a Bayesian method for feature selection has been applied to genes taken from *Arabidopsis thaliana* (thale cress) [15]. In this case, the overall accuracy obtained was around 70% of the genes considered, but the analysis was performed only on two particular condition–specific expression sets.

Table 1
The seven yeast gene classes used for the pairwise cluster classification, with number of genes, and functional enrichment of the genes contained in each.

| Cluster No. | # of genes | Functional class |
|---|---|---|
| 1 | 138 | ribosome biogenesis |
| 2 | 123 | none |
| 3 | 115 | none |
| 4 | 114 | rRNA transcription |
| 5 | 89 | C-compound metabolism |
| 6 | 86 | aminoacyl-tRNA-synthetases |
| 7 | 84 | stress response |

## 5 Results

### *5.1 Experimental setup*

Classes of co-expressed genes in *S. cerevisiae* (yeast) have been obtained from the clustering results reported in [12]: 49 clusters of genes with similar expression profile have been detected, using a modified version of the k-means algorithm applied to DNA microarray experiments relative to environmental stresses [16] and cell cycle [17], for 255 total conditions. From the original 49 we excluded eight clusters with a too low number of examples, in order to obtain more reliable supervised predictions. As a consequence, for our experiments we used only the 41 clusters with the highest number of examples. Clusters are numbered according to their size (from the largest to the smallest), and we denote them with $C_1, C_2, \ldots, C_{41}$.

For each gene, we retrieved from the *S. cerevisiae* promoter database (SCPD, [20]) the 800 base pairs upstream of the start codon (as in [12]). Motif extraction was performed by the Weeder algorithm using motifs of length 6 and 8, allowing one substitution (mismatch) in their occurrences.

We considered two types of experiments:
(a) pairwise classification of a single cluster against another single cluster;
(b) classification of a single cluster against a random combination of examples chosen from the remaining clusters, that is, genes picked at random.

More precisely, for type (a) experiments, from the set $U = \{C_1, C_2, \ldots, C_{41}\}$ of all the clusters used in our experiments, we selected the set $S = \{C_1, C_2, \ldots, C_7\}$ of the first seven clusters (the clusters containing more examples, summarized in Table 1 together with their functional enrichment, that is, the gene func-

Table 2
Summary of the classification results.

| **Pairwise classification** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Kernel/mers** | $\mathbf{C_1}$ | $\mathbf{C_2}$ | $\mathbf{C_3}$ | $\mathbf{C_4}$ | $\mathbf{C_5}$ | $\mathbf{C_6}$ | $\mathbf{C_7}$ | **Mean** |
| Sigmoid sixmers | 0.7646 | 0.6897 | 0.7205 | 0.7427 | 0.7572 | 0.6833 | 0.7354 | **0.7276** |
| Sigmoid eightmers | 0.7797 | 0.6778 | 0.6927 | 0.7686 | 0.7377 | 0.6862 | 0.7066 | **0.7213** |
| **One vs random classification** | | | | | | | | |
| Sigmoid sixmers | | | **0.6644** | Sigmoid eightmers | | | | **0.6509** |

tion that is significantly over–represented in the cluster). Then for each cluster $C \in S$ we performed a pairwise classification against each $C' \in U$, such that $C' \neq C$, for a total of $7 \times 40 = 280$ pairwise classification tasks.

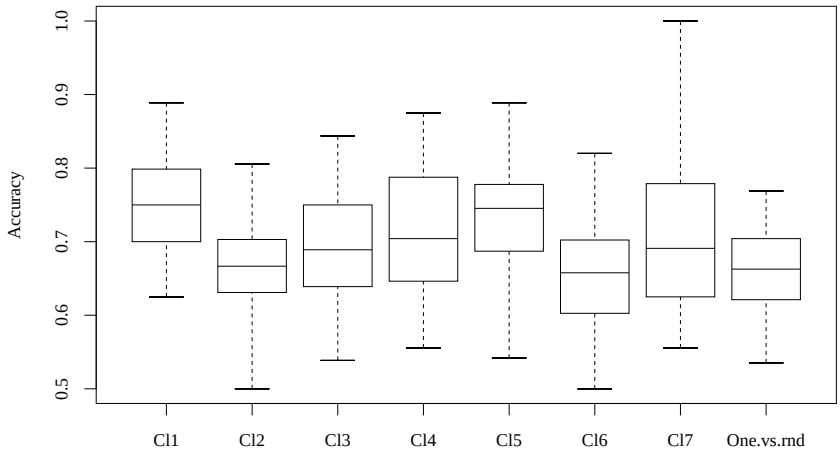
For type (b) experiments we considered a classification of each of the clusters $C \in U$ against a random choice of examples belonging to the remaining 40 clusters. That is, for each $C \in U$ we built up the corresponding negative set of examples by randomly choosing examples from the set $R = \bigcup_{C' \in U, C' \neq C} C'$, for a total of 41 pairwise classification tasks.

In these dichotomic classification tasks we balanced the number of the examples of the two classes: about the same number of examples has been chosen for both.

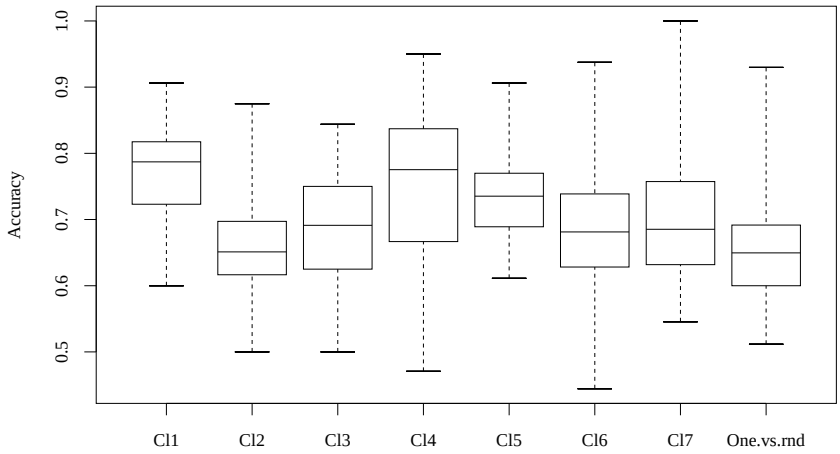
To estimate the classification performance of SVMs we applied a multiple hold–out technique: the overall data have been randomly split 5 times in a training set (70% of examples) and a test set (30%) and the results on the test set have been averaged. We carefully considered the selection bias problem [18]: for each training set we extracted from the promoters a subset of motifs using a t-test on the motif scores computed by Weeder (see Sect. 3.1), and we used the same subset of motifs to estimate the generalization error on the test set. Note that in this way it is likely that for each train/test split of the data we select a slightly different subset of motifs. For each training set, we selected motifs at 0.005 significance level, according to the two-sample t-test; in case the test selected less than 100 motifs, we used the 100 top ranked motifs according to the Welch t–statistic.

### *5.2 Classification of co-expressed genes*

At first we tried to apply linear methods to separate different classes of co-expressed genes. In particular we used perceptrons and linear SVMs, with quite unsatisfactory results. The average accuracy of the perceptron was around 55%, while linear SVMs reached 65%; thus, in order to obtain a higher classification accuracy we employed non–linear methods.



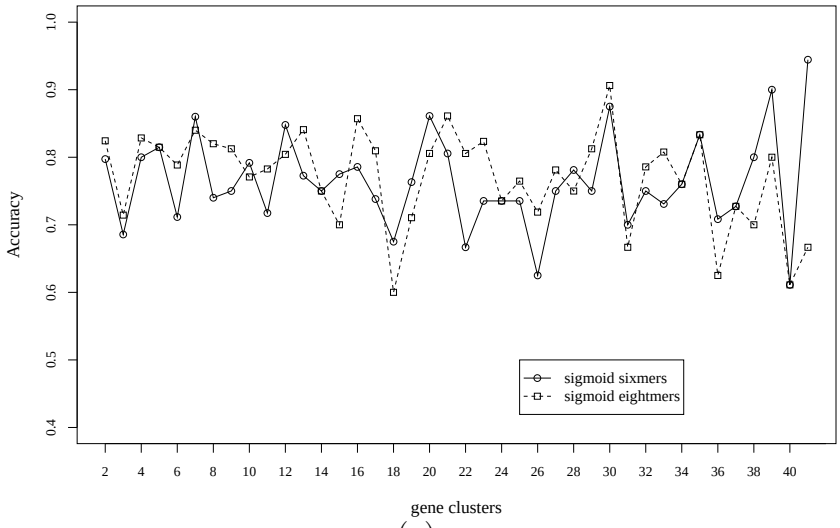

(a)



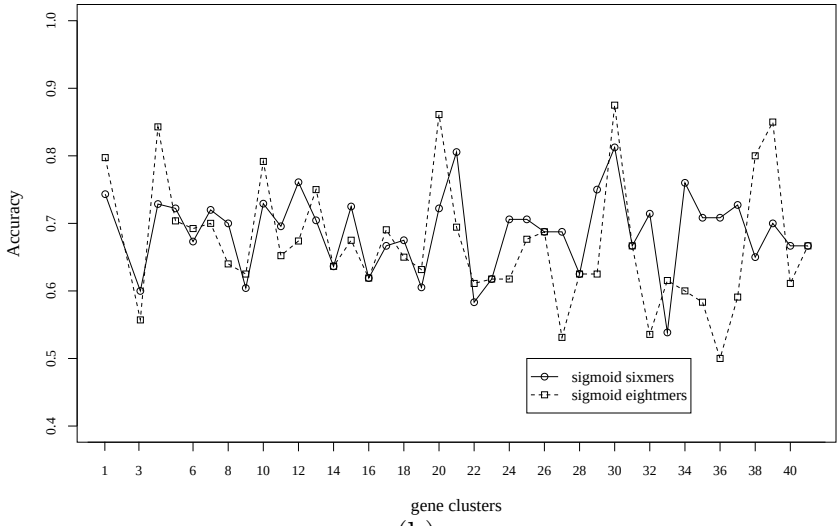

(b)

Fig. 3. Boxplots of the pairwise classification of the first 7 clusters against all the others; the last boxplot represents the one vs random subsets of the other clusters accuracy distributions (a) Sigmoid kernels applied to sixmers (b) Sigmoid kernel applied to eightmers.

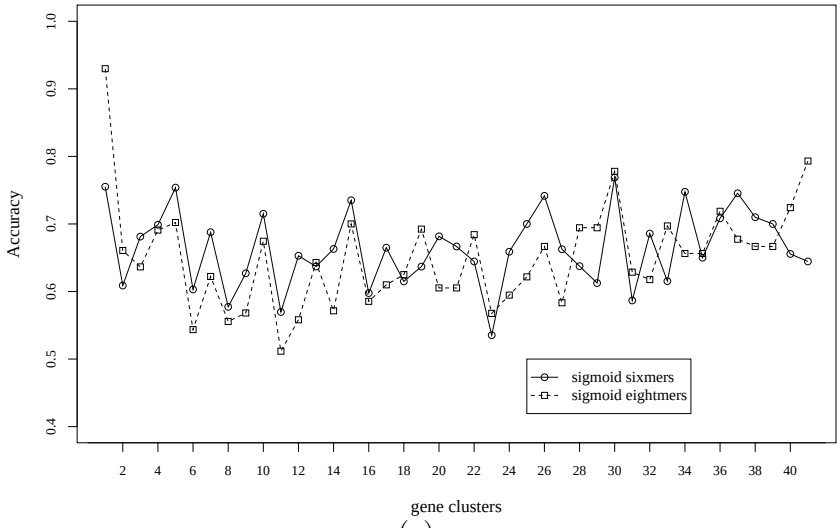
As noted in Sect. 3.2, we used gaussian and sigmoid kernels. We performed a tuning of the parameters of the SVMs, considering $C$ values between $10^{-3}$ and $10^3$, and varying the $\sigma$ parameter of the gaussian kernel between $10^{-5}$ and $10^5$, and the $c_1$ parameter of the sigmoid kernel between $10^{-5}$ and $10^5$ as well (see Sect. 3.2). In any case we performed only a coarse tuning of the parameters, because our main aim was to understand whether gene expression could be predicted from local sequence data in gene promoters.




(a)




(b)




(c)

Fig. 4. SVMs with sigmoid kernel classification accuracies; results obtained with sixmers and eightmers are compared. (a) Pairwise classification of the first cluster vs each one of the other clusters (b) Pairwise classification of the second cluster vs each one of the other clusters (c) Classification of each cluster vs. a random set of genes selected from the others.

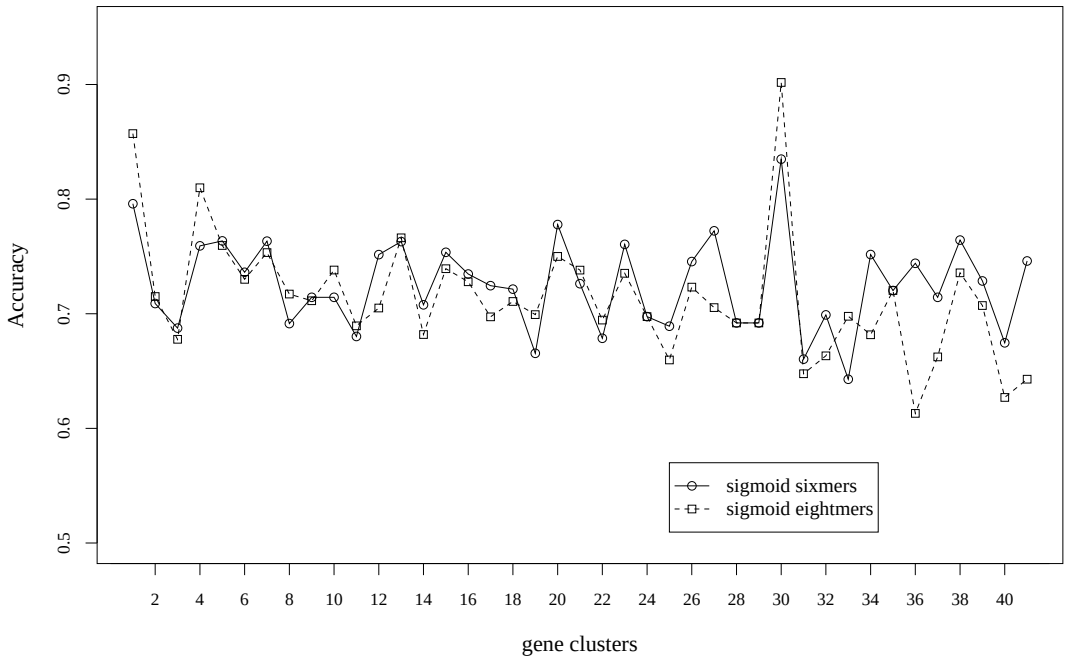



Fig. 5. Mean pairwise classification accuracies of each cluster vs each of the first seven clusters. SVM sigmoid kernels results obtained with sixmers and eightmers are compared.

Table 2 summarizes the overall mean results of the pairwise classification. The first column represents the kernel type/motif length (six or eight) used in the experiment. The columns from $\mathbf{C_1}$ to $\mathbf{C_7}$ represent the average accuracy of each of the clusters $C \in S$ vs. each of the other clusters $C' \in U$. The last column simply reports the mean of each row, that is, the "mean of the means" with respect to the clusters $C \in S$. The "mean of the means" column shows that for both sixmers and eightmers we obtain an accuracy above 0.70. In particular, concerning single clusters, the percentage of genes in test sets correctly classified ranged from 68% to 76%, thus showing the robustness of the method regardless of the clusters selected for the comparison. The gaussian kernel yielded a slightly lower accuracy (about 2% less on the average) both with sixmers and eightmers (data not shown). The accuracy we obtained is similar to the result reported in [12], although a direct comparison is not possible. First of all, we compared pairs of classes, while in [12] a multi–class classification was performed; also, we did not resort to known TFBSs instances.

The last row of the table instead summarizes the results of the classification of each of the 41 clusters versus a random subset of examples drawn at random from the other classes (experiment (b)). Also in this case the accuracy is significantly larger than 0.5, with an average of about 66.5%. The lower accuracy w.r.t. the previous test can be explained by the fact that, by choosing genes at random, we had no guarantee of picking genes with a completely different expression profile, that is, regulated by different sets of motifs as in the pairwise cluster comparison. Again, the gaussian kernel reached a slightly lower

accuracy level.

Figure 3 shows the distribution of the accuracy summarized in Table 2. The box–plots represent the distribution of the accuracy of the pairwise and one–versus–random classification tasks. Independently of the length of the motifs, the distributions of the accuracies are significantly higher than 0.5 in all the clusters. The variations can be also due to the fact that some of the clusters are less "well defined", and thus the corresponding genes have an expression profile similar to other genes belonging to other clusters.

Figure 4 details the results w.r.t. the pairwise classification of cluster 1 (Fig. 4a) and cluster 2 (Fig 4b) against each one of the other 40 clusters, as well as the classification of each of the 41 clusters against a random gene set (Fig. 4c).

The results show that there is no definite preference on the motif size used for the classification: better results are achieved either by six- or eightmers for different pairs of clusters.

Anyway, by looking at Fig. 5, showing the mean pairwise classification accuracy of each cluster vs. each of the first seven, we can see that there is no significant difference on the average accuracy obtained by six- or eightmers. The only visible difference is on the worse performance of eightmers in the last clusters, that can be due to the small size of the clusters themselves (15–20 genes in each), that in turn makes possible the extraction of a too small number of significantly conserved motifs of size eight, making the classification task harder. This fact is mirrored by the greater accuracy yielded by eightmers in the first (largest) classes (more than 100 genes in each). All in all, both Fig. 4 and 5 show that the overall accuracy is quite well balanced by the accuracy obtained in each comparison, with no significant oscillations. Also, the sensitivity and specificity values obtained in all the experiments are balanced and consistent with accuracy levels, while linear classifiers tended to generate unbalanced errors with significant differences between sensitivity and specificity (data not shown).

## 6 Conclusions and Further Work

The results of the tests we performed can be considered quite satisfying. The classification accuracy is comparable with other state of the art studies based on the same principle, that report an accuracy rate around 70%, showing all the potential of an heterogeneous approach merging sequence analysis with microarray clustering, and combinatorial algorithms with statistical and machine learning approaches.

An important difference between our method and recent literature is that, even if the overall accuracy is comparable, in our work we did not resort to descriptors of known TFBSs as in most of the other approaches. This choice is due to the fact that our main goal was to assess the feasibility of a complete ab initio fully automated method, without the introduction of any prior knowledge of the factors involved in the regulation of the genes. Including this information, we can reasonably expect could further improve the reliability of our method.

Other improvements could be obtained by a more fine–grained tuning of the SVM parameters, which in our experiments played an essential role in the performance of the classifier: small variations in fact led to significant improvements. Also, we employed fixed significance thresholds for the motif selection step: motifs scored by Weeder could be submitted to a further procedure explicitly aimed at selecting the optimal subset yielding the best discrimination between two classes of genes, in turn providing another measure to assess the significance of conserved motifs, one of the most challenging open problems of motif discovery. Also, we did not consider explicit information about position, orientation and correlations between motifs in the sequences analyzed, all factors that have been proven to play significant roles in the regulation of transcription, especially in metazoa. As a future work we aim to investigate if correlations among different motifs could be significant for identifying classes of co-regulated genes. To this end we could apply wrapper methods such as forward and backward feature selection methods or floating search methods with backtracking [19] that are able to take into account interactions between motifs in the feature selection process.

Moreover, the presence of conserved TFBSs in a promoter sequence does not guarantee the transcription of the corresponding gene, since also the TF binding the site has to be present. This kind of information is anyway very hard to gather, since it involves knowledge about the association between the DNA binding specificity of the TFs and the conserved motifs found in the sequences. Another important issue that has to be considered is that clusters of co-expressed genes (like the ones we employed in our experiments) are obtained by the application of clustering algorithms: thus, some genes might be "mis–labelled" or have low correlation in expression with the other genes assigned to the same cluster, and hence very hardly co-regulated and likely to undergo a different regulation process.

From an experimental point of view, the next logical step in our work is to apply our technique to other organisms, like human, where the regulation of transcription is governed in a more complex way, involving regions outside the core promoter of the genes. Recent research, however, has pointed to the opposite direction, showing for example that promoter sequences in human and mouse genes differ according to the tissue–specificity of the gene itself [14].

Applying our method also to case studies like these will perhaps help to shed further light on this issue, that remains one of the most relevant and studied in bioinformatics and molecular biology.

## Acknowledgments

We would like to thank the anonymous reviewers for their comments and suggestions.

## References

[1] H. Pearson, What is a gene, Nature 441 (25) (2006) 399–401.

[2] G. Pavesi, G. Mauri, G. Pesole, In silico representation and discovery of transcription factor binding sites, Brief Bioinform 5 (3) (2004) 217–36.

[3] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, Z. Zhu, Assessing computational tools for the discovery of transcription factor binding sites, Nat Biotechnol 23 (1) (2005) 137–44.

[4] J. De Risi, V. Iyer, P. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, Science 278 (1997) 680–686.

[5] M. Eisen, P. Spellman, P. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, PNAS 95 (25) (1998) 14863–14868.

[6] P. Gasch, M. Eisen, Exploring the conditional regulation of yeast gene expression through fuzzy k-means clustering, Genome Biology 3 (11).

[7] A. D. Smith, P. Sumazin, M. Q. Zhang, Identifying tissue-selective transcription factor binding sites in vertebrate promoters, Proc Natl Acad Sci U S A 102 (5) (2005) 1560–1565.

[8] G. Pavesi, P. Mereghetti, G. Mauri, G. Pesole, Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes, Nucleic Acids Res 32 (Web Server issue) (2004) W199–203.

[9] N. Li, M. Tompa, Analysis of computational approaches for motif discovery, Algorithms Mol Biol 1 (1) (2006) 8.

[10] V. N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[11] J. Mercer, Function of positive and negative type and their connection with the theory of integral equations, Philos. Trans. Roy. Soc. London 209 (1909) 415–446.

[12] M. Beer, S. Tavazoie, Predicting gene expression from sequence, Cell 117 (2004) 185–198.

[13] R. Sharan, E. W. Myers, A motif-based framework for recognizing sequence families, Bioinformatics 21 Suppl 1 (2005) i387–i393.

[14] A.D.Smith, P.Sumazin P, Z.Xuan, M.Q.Zhang, DNA motifs in human and mouse proximal promoters predict tissue-specific expression, Proc Natl Acad Sci USA 103 (2006) 6275–6280.

[15] Y. Li, K. Lee, S. Walsh, C. Smith, S. Hadingham, K. Sorefan, G. Cawley, M. Bevan, Establishing glucose- and ABA-regulated transcription networks in Arabidopsis by microarray analysis and promoter classification using a Relevance Vector Machine, Genome Research 16 (2006) 414–427.

[16] P. Gasch, et al., Genomic expression programs in the response of yeast cells to environmental changes, Mol.Biol.Cell 11 (2000) 4241–4257.

[17] P. Spellman, et al., Comprehensive identification of cell cycle-regulated genes of the yeast saccharomices cerevisiae by microarray hybridization, Mol. Biol. Cell 9 (1998) 3273–3297.

[18] C. Ambroise, G. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data, PNAS 99 (10) (2002) 6562–6566.

[19] P. Somol, P. Pudil, J. Novovicova, P. Paclik, Adaptive floating search methods in feature selection, Pattern Recognition Letters, 20 (11/13) (1999) 1157-1163.

[20] J. Zhu, M. Q. Zhang, SCPD: a promoter database of the yeast saccharomyces cerevisiae, Bioinformatics 15 (7-8) (1999) 607–11.