Methodology article

Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes

Michele Caselle¹, Ferdinando Di Cunto² and Paolo Provero*³, ¹

Address: ¹Dipartimento di Fisica Teorica, Università di Torino, and INFN, Sezione di Torino, Torino, Italy, ²Dipartimento di Genetica, Biologia e Biochimica, Università di Torino, Torino, Italy and ³Dipartimento di Scienze e Tecnologie Avanzate, Università del Piemonte Orientale, Alessandria, Italy

E-mail: Michele Caselle - caselle@to.infn.it; Ferdinando Di Cunto - ferdinando.dicunto@unito.it; Paolo Provero* - provero@to.infn.it

*Corresponding author

Abstract

Background: Gene regulation in eukaryotes is mainly effected through transcription factors binding to rather short recognition motifs generally located upstream of the coding region. We present a novel computational method to identify regulatory elements in the upstream region of eukaryotic genes. The genes are grouped in sets sharing an overrepresented short motif in their upstream sequence. For each set, the average expression level from a microarray experiment is determined: If this level is significantly higher or lower than the average taken over the whole genome, then the overerpresented motif shared by the genes in the set is likely to play a role in their regulation.

Results: The method was tested by applying it to the genome of Saccharomyces cerevisiae, using the publicly available results of a DNA microarray experiment, in which expression levels for virtually all the genes were measured during the diauxic shift from fermentation to respiration. Several known motifs were correctly identified, and a new candidate regulatory sequence was determined.

Conclusions: We have described and successfully tested a simple computational method to identify upstream motifs relevant to gene regulation in eukaryotes by studying the statistical correlation between overrepresented upstream motifs and gene expression levels.

Introduction

One of the biggest challenges of modern genetics is to extract biologically meaningful information from the huge mass of raw data that is becoming available. In particular, the availability of complete genome sequences on one hand, and of genome-wide microarray data on the other, provide invaluable tools to elucidate the mechanisms underlying transcriptional regulation. The sheer amount of available data and the complexity of the mechanisms at work require the development of specific data analysis techniques to identify statistical patterns and regularities, that can then be the subject of experimental investigation.

The regulation of gene expression in eukaryotes is known to be mainly effected through transcription factors binding to rather short recognition motifs generally located upstream of the coding region. One of the main problems in studying regulation of gene expression is to identify the
motifs that have transcriptional meaning, and the genes each motif regulates.

The usual approach to this kind of analysis begins by identifying groups of co-regulated genes, for example by applying clustering techniques to the expression profiles obtained from microarray experiments. One then studies the upstream sequences of a set of coregulated genes looking for shared motifs. Examples of this approach as applied to S. cerevisiae are Refs. [1,2,4].

In this paper we suggest an alternative method which somehow follows the inverse route: genes are grouped into (non-disjoint) sets, each set being characterized by a short motif which is overrepresented in the upstream sequence. For each set, the average expression is computed for a certain microarray experiment, and compared to the genome-wide average expression from the same experiment. If a statistically significant difference is found, then the motif that defines the set of genes is a candidate regulatory sequence. The rationale for looking for overrepresented motifs is that, in many instances, regulatory motifs are known to appear repeated many times within a relatively short upstream sequence [2,3], so that the number of repetitions turns out to be much bigger than what would be expected from chance alone.

A somehow related approach, which does not require any previous grouping of genes based on their expression profiles, was presented in Ref. [5], where the effect of upstream motifs on gene expression levels is modeled by a sum of activating and inhibitory terms. Experimental expression levels are then fitted to the model, and statistically significant motifs are identified. Our approach differs in the importance given to overrepresented motifs, thus considering activation and inhibition as an effect that depends on a threshold number of repetitions of a motif rather than on additive contributions from all motifs. Clearly the two mechanisms are far from being mutually exclusive, therefore we expect the candidate regulatory sites found with the two methods to significantly overlap.

However it is important to notice that the kind of statistical correlation between upstream motifs and expression that our algorithm identifies does not depend on any special assumption on the functional dependence of expression levels on the number of motif repetitions, as long as this dependence is strong enough to provide a significant deviation from the average expression when enough copies of the motif are present. A comparison of our results with those obtained in Ref. [5] is provided in the "Results and discussion" section.

The method

In general the motifs with known regulatory function are not identified with a fixed nucleotide sequence, but rather with sequences where substitutions are allowed, or spaced dyads of fixed sequences, etc. However in this study, in order to test the method while keeping the technical complications to a minimum, we will limit ourselves to fixed short nucleotide sequences, that we call words. While previous studies (see e.g [2]) show that even this simple analysis can give interesting results, the method we present can easily be generalized to include variable sequences and other more complicated patterns.

The computational method we propose has two main steps: first the open reading frames (ORFs) of an eukaryote genome are grouped in (overlapping) sets based on words that are overrepresented in their upstream region, compared to their frequencies in the reference sample made of all the upstream regions of the whole genome. Each set is labelled by a word. Then for each of these sets the average expression in one or more microarray experiments are compared to the genome-wide average: if a statistically significant difference is found, the word that labels the set is a candidate regulatory site for the genes in the set, either enhancing or inhibiting their expression.

It is worth stressing that the grouping of the genes into sets depends only on the upstream sequences and not on the microarray experiment considered: it needs to be done only once for each organism, and can then be used to analyse an arbitrary number of microarray experiments. It is precisely this fact that should allow the extension of the method to patterns more complex than fixed sequences, while keeping the required computational resources within reasonable limits.

Constructing the sets

We consider the upstream region of each open reading frame (ORF), and we fix the maximum length $K$ of the upstream sequence to be considered. The choice of $K$ depends on the typical location of most regulatory sites: in general $K$ is a number between several hundred and a few thousand. For each ORF $g$, the actual length of the sequence we consider is $K_g$ defined as the minimum between $K$ and the available upstream sequence before the coding region of the previous gene.

For each word $w$ of length $l$ ($6 \leq l \leq 8$ in this study), and for each ORF $g$ we compute the number $n_g(w)$ of occurrences of $w$ in the upstream region of $g$. Non palindromic words are counted on both strands: therefore we define the effective number of occurrences $n_g(w)$ as
where $\bar{w}$ is the reverse complement of $w$.

We define the global frequency $p(w)$ of each word $w$ as

$$p(w) = \frac{\sum_{g} n_g(w)}{\sum_{g} L_g(w)}$$

where, in order to count correctly the available space for palindromic and non palindromic words,

$$L_g(w) = 2 \left( K_g - l + 1 \right) \quad \text{if} \quad w \neq \bar{w}$$

$$L_g(w) = 2 \left( K_g - l + 1 \right) \quad \text{if} \quad w = \bar{w}$$

$p(w)$ is therefore the frequency with which the word $w$ appears in the upstream regions of the whole genome: it is the “background frequency” against which occurrences in the upstream regions of the individual genes are compared to determine which words are overrepresented.

For each ORF $g$ and each word $w$ we compute the probability $b_g(w)$ of finding $n_g(w)$ or more occurrences of $w$ based on the global frequency $p(w)$:

$$b_g(w) = \sum_{n=n_g(w)}^{L_g(w)} \binom{L_g(w)}{n} p(w)^n \left[ 1 - p(w) \right]^{L_g(w)-n}$$

We define a maximum probability $P$, depending in general on the length $l$ of the words under consideration, and consider, for each $w$, the set

$$S(w) = \{ g : b_g(w) < P \}$$

of the ORFs in which the word $w$ is overrepresented compared to the frequency of $w$ in the upstream regions of the whole genome. That is, $w$ is considered overrepresented in the upstream region of $g$ if the probability of finding $n_g(w)$ or more instances of $w$ based on the global frequency is less than $P$.

This completes the construction of the sets $S(w)$. Two free parameters have to be fixed: the length $K$ of the upstream region to be considered and the probability cutoff $P$ for each length $l$ of words considered. A result in Ref. [2] suggests suitable choices of these two numbers: the authors list the 34 ORFs of $S.\ cer\visea$ that have 3 or more occurrences of the word GATAAG in their 500 bp upstream region. 23 out of these 34 ORFs correspond to a gene with known function, and 20 out of these 23 are regulated by nitrogen. This result suggests to choose $K = 500$ for the upstream length, and a value of the probability cutoff such that three or more instances of GATAAG in the 500 bp upstream region of an ORF are considered significant. Any choice of $P$ between 0.018 and 0.1 would satisfy this criterion, and we chose $P = 0.02$. Tentatively, we kept the same value of $P$ for all values of $l$. With this choice, the number of instances of a word that are necessary to be considered overrepresented in a 500 bp upstream sequence can be as high as six for common 6-letter words and as low as one for rare 8-letter words. In particular, our set $S(\text{GATAAG})$ almost coincides with the one discussed in [2]. However the word GATAAG will not turn out to be significant in our study.

As noted above, it would be natural to make the probability cutoff $P$ depend on the word length, simply because the number of possible words increases with their length: For example one could take the cutoff for each word length to be inversely proportional to the number of independent words of such length. However it turns out that this procedure tends to construct sets that are less significant when tested for correlation with expression. Therefore we chose to fix the cutoff at 0.02 for all word lengths. It is important to keep in mind that no statistical significance whatsoever is attributed to the sets per se: The only sets that are retained at the end of the analysis are the ones that show significant correlation with expression. Therefore the choice of the cutoff in the construction of the sets can be based on such a pragmatic approach without jeopardizing the statistical relevance of the final result.

---

| $i$ | $N(i)$ | $R(i)$ | $\sigma(i)$ |
|-----|--------|--------|-------------|
| 1   | 6082   | -0.0888| 0.2509      |
| 2   | 6054   | -0.0378| 0.2801      |
| 3   | 6020   | 0.1132 | 0.3152      |
| 4   | 6071   | -0.1957| 0.3433      |
| 5   | 6058   | -0.2423| 0.3890      |
| 6   | 6084   | 0.09244| 0.8226      |
| 7   | 6021   | -0.2028| 0.8886      |

Table 1: Number of data, average and standard deviation for the 7 time-points.
Studying the average expression level in each set

The second step of our procedure consists in studying, for each set $S(w)$ defined as above, the expression profiles of the ORFs belonging to $S(w)$ in DNA microarray experiments. The idea is that if the average expression profile in the set $S(w)$ for a certain experiment is significantly different from the average expression for the same experiment computed on the whole genome, then it is likely that some of the ORFs in $S(w)$ are coregulated and that the word $w$ is a binding site for the common regulating factor.

To look for such instances we consider the gene expression profiles during the diauxic shift, i.e. the metabolic shift from fermentation to respiration, as measured with DNA microarrays techniques in Ref. [1]. In the experiment gene expression levels were measured for virtually all the genes of $S. Cerevisiae$ at seven time-points while such metabolic shift took place. The experimental results are publicly available from the web supplement to Ref. [1].

We considered each time-point as a single experiment, and for each gene $g$ we defined the quantity $r_g(i)$ ($1 = 1, ..., 7$) as the log$_2$ of the ratio between the mRNA levels for the gene $g$ at time-point $i$ and the initial mRNA level. Therefore e.g. $r_g(i) = 1$ means a two-fold increase in expression at timepoint $i$ compared to initial expression.

For each time-point $i$ we computed the genome-wide average expression $R(i)$ and its standard deviation $\sigma(i)$. These are reported in Tab. 1, where $N(i)$ is the number of genes with available expression value for each timepoint. Then for each word $w$ we compute the average expression in the subset of $S_w$ given by the genes for which an experimental result at timepoint $i$ is available, and the difference

$$\Delta R_w(i) = R_w(i) - R(i)$$

$\Delta R_w(i)$ is the discrepancy between the genome-wide average expression at time-point $i$ and the average expression at the same time-point of the ORFs that share an abundance of the word $w$ in their upstream region. A significance index $\text{sig}(i, w)$ is defined as

$$\text{sig}(i, w) = \frac{\Delta R_w(i)}{\sigma(i) \sqrt{N(i, w)}}$$

and the word $w$ is considered significantly correlated with expression at time point $i$ if

$$|\text{sig}(i, w)| > \Lambda$$

In this work we chose $\Lambda = 6$: this means that we consider meaningful a deviation of $R_w(i)$ by six s.d.’s from its expected value. The sign of $\text{sig}(i, w)$ indicates whether $w$ acts as an enhancer or an inhibitor of gene expression.

| Table 2: Significant words related to the PAC motif. |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| word            | genes | timepoints | | score |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| GATGAG          | 24    | -        | -        | -4.70 | -        | -        | -        | 1.00 |
| GATGAGAT        | 35    | -        | -        | -8.20 | -6.26 | -6.18 | -7.86 | 0.94 |
| GATGAGA         | 26    | -        | -        | -7.06 | -4.64 | 0.93 |
| GAGATGAG        | 36    | -        | -        | -6.96 | -4.50 | 0.92 |
| AGATGAG         | 33    | -        | -        | -6.17 | -4.44 | 0.91 |
| GAGATGA         | 42    | -        | -        | -6.20 | -4.33 | 0.83 |
| ATGAGATG        | 32    | -        | -        | -6.96 | -4.33 | 0.80 |
| GAGATG          | 31    | -        | -        | -6.42 | -4.25 | 0.75 |
| TGAGATG         | 47    | -        | -        | -6.26 | -4.10 | 0.70 |
Results and discussion

We found a total of 29 words of length between 6 and 8 above our significance threshold $|\text{sig}| > 6$. Most of them are related to known regulatory motifs; two words turned out to be false positives due to the presence, in their sets, of families of identical ORF's. Finally, one word does not match any known motif and is a candidate new binding site.

The comparison between our significant words and known motifs was performed using the database of regulatory motifs made publicly available by the authors of Ref. [6], and the CompareACE software [7] available from the same web source. This package allowed us to compute the Pearson correlation coefficient of the best alignment between each of our significant words and each known regulatory motif (expressed as a set of nucleotide frequencies).

We used the following criterion to associate our significant words to known motifs: a motif is considered as identified if at least one significant word scores better than 0.8 when compared to it. A probability value for this choice of the cutoff can be estimated to be a few percent: out of all the 2080 independent 6-letter words, 66 (that is 3.17%) score better than 0.8 with at least one motif. For 7- and 8-letter words we have respectively 2.21% and 1.51%. Once a motif has been identified, all words which score best with the motif are attributed to it, independently of the score, provided their expression pattern is consistent with the word(s) scoring better than 0.8.

PAC and RRPE motifs

Nine significant words can be associated to the PAC motif [8,4,7], all of them with rather high scores. They are shown in Tab. 2, where, as in all the following tables, significativity indices are shown only for those timepoints where they exceed our threshold $|\text{sig}| > 6$. Given the perfect alignment of these words, it is not surprising that these sets largely overlap each other: The union of all the nine sets contains a total of 96 genes. As an example, in Fig. 1 we show the average expression for the genes associated with the word GATGAG as a function of the time, compared to the average expression computed over the whole genome. Fig. 2 shows the significance index for the same set. The genes in this set are shown in Tab. 3 together with their expression profiles.

Two words can be associated with confidence to the motif RRPE [4,7], and are shown in Tab. 4. The union of the two sets contains 76 genes. We see that genes containing the motifs PAC and RRPE are repressed at the late stage of the diauxic shift compared to the early stages. This result is in agreement with the expression coherence score data available from the web supplement to Ref. [6]: There one can see that (1) of all known regulatory motifs, PAC and RRPE show the highest expression coherence for the diauxic shift and (2) vice versa, of the eight experimental conditions considered in Ref. [6], the diauxic shift is the one in which both the PAC and RRPE motif show the highest expression coherence score.

STRE and MIG1 motifs

A total of ten significant words can be associated to the motifs STRE [9,10] and MIG1 [11,12]. It is well known
that these play an important role in glucose repression (see e.g. [1,13] and references therein). Most of these words show comparable scores for the two motifs (due to their similarity) so we decided to show them together in Tab. 5 which shows the two scores for each word. A total of 212 genes belong to the union of all these sets.

**Table 3**: The ORFs in the set S(GATGAG) with their expression profiles.

| ORF     | gene | 1    | 2    | 3    | 4    | 5    | 6    | 7    |
|---------|------|------|------|------|------|------|------|------|
| YBL054W |      | 0.21 | -0.01| -0.18| -1.56| -1.25| -0.79| -1.47|
| YCL059C | KRR1 | 0.36 | 0.06 | 0.45 | -0.69| -0.71| -0.34| -1.69|
| YDL063C |      | -0.03| -0.03| -0.27| -0.92| -1.06| -1.51| -2.06|
| YDL153C | SAS10| 0.41 | 0.19 | 0.36 | -0.76| -0.97| -1.43| -1.79|
| YDR365C |      | 0.03 | 0.06 | 0.21 | -0.38| -0.62| -1.64| -1.94|
| YGR022C |      | -0.17| -0.06| 0.14 | 0.04 | -0.15| 0.54 | 0.86 |
| YGR102C |      | -0.23| -0.23| -0.07| -0.32| 0.03 | 1.43 | 0.84 |
| YGR103W | NOP7 | 0.15 | -0.06| 0.32 | -0.92| -1.09| -1.64| -2.56|
| YGR128C |      | 0.30 | 0.26 | 0.38 | -0.81| -0.76| -0.89| -1.47|
| YGR129W | SYF2 | -0.18| -0.54| 0.11 | -0.12| -0.23| 0.74 | 0.14 |
| YGR145W |      | 0.00 | -0.23| 0.25 | -0.92| -1.09| -1.69| -2.18|
| YJL033W | HCA4 | -0.06| 0.01 | 0.21 | -0.94| -0.36| -0.67| -0.62|
| YKL078W |      | -0.04| -0.01| 0.04 | -1.12| -0.97| -0.71| -1.89|
| YKL172W | EBP2 | 0.12 | 0.21 | 0.30 | -0.74| -0.56| -0.42| -1.40|
| YLR276C | DBP9 | 0.03 | 0.14 | 0.32 | -0.62| -0.86| -0.67| -1.64|
| YLR401C |      | -0.06| -0.07| 0.07 | -0.71| -0.71| -0.84| -1.03|
| YLR402W |      | -0.18| -0.23| -0.30| -0.47| -0.51| -0.20| -0.27|
| YML123C | PHO84| 0.50 | 0.50 | 0.54 | -0.56| -0.67| -2.32| -1.69|
| YNL061W | NOP2 | -0.03| -0.51| -0.42| -1.29| -1.36| -2.25| 0.01 |
| YNL062C | GCD10| -0.10| 0.00 | 0.01 | -0.47| -0.64| -1.12| -1.06|
| YOL141W | PPM2 | -0.10| 0.01 | 0.24 | -0.84| -0.54| 0.04 | -0.20|
| YPL068C |      | -0.60| -0.10| -0.18| -0.84| -1.09| 0.08 | -0.89|
| YPR112C | MRD1 | -0.17| -0.23| -0.17| -0.54| -0.62| -1.12| -1.51|
| YPR113W | PIS1 | -0.04| 0.00 | 0.62 | 0.52 | 0.56 | 1.12 | -1.03|

| set average | 0.005 | -0.036| 0.124 | -0.666| -0.676| -0.679| -1.106|
| genome average | -0.089| -0.038| 0.113 | -0.196| -0.242| 0.092 | -0.203|
| significance | 1.83 | 0.03 | 0.17 | -6.71 | -5.47 | -4.60 | -4.98 |

**Table 4**: Significant words related to the RRPE motif.

| word     | genes | 1    | 2    | 3    | 4    | 5    | 6    | 7    | score  |
|----------|-------|------|------|------|------|------|------|------|-------|
| AAAATTTT | 50    | -    | -    | -    | -    | -    | -7.90| -8.58| 0.91  |
| AAAATTTT | 62    | -    | -    | -6.59| -    | -8.73| -10.26| 0.89  |

**The UME6 motif**

Two words are associated to the known UME6 motif, a.k.a. URS1 [14,15], known to be a pleiotropic regulator implicated in glucose repression [16]. They are shown in Tab. 6. The two sets do not overlap, so that a total of 56 genes are associated to this motif.
Other significant words

Three words, shown in Tab. 7, are of uncertain status: for the first one, the set S(ACTTTC) contains only 2 genes, making the statistical significance of the result questionable. The word CCCCTGAA scores best with the PDR motif (0.58): given the low significance of this score, and the fact that PDR does not seem to be relevant for any other word, this is most likely accidental. The word should probably be considered as belonging to the STRE/MIG1 motif (the scores are STRE: 0.46, MIG1: 0.49). Finally the word GCCCCTGA scores best with UME6 (0.55), but its expression pattern is more similar to the STRE/MIG1 motifs (scores: STRE:0.44, MIG1: 0.46).

False positives due to families of identical or nearly identical ORF’s

The genome of *S. cerevisiae* contains a few families of genes whose coding and upstream regions are identical or nearly identical. Consider for example the COS1 gene (YNL336W): the seven genes COS2-COS8 have both cod-
ing sequence and 500 kb upstream sequence coinciding better than 80% with the COS1 sequence. Therefore if the upstream sequence of COS1 contains over-represented words, they will likely appear in all of the upstream regions. On the other hand, the expression profiles of all the genes in the family will be the same when measured by a microarray experiment, simply because the experimental apparatus cannot distinguish between the mRNA produced by the various members of the family, due to cross-hybridization between their mRNA. Therefore all of the genes of the family are likely to occur in the sets of the words that are overrepresented in their upstream region, and even a small deviation from the genome-averaged expression acquires a statistical significance.

We found two instances of this in our analysis: the words GACGTAGC and GGTCGCAC appear to be associated to significant enhancement of the corresponding sets of genes at late timepoints in the diauxic shift: however the two sets contain respectively seven out of eight and all of the COS1-COS8 genes. Since the COS genes are mildly overexpressed, this creates a false statistical significance. When one corrects for this, by keeping only one representative of the family, the statistical significance of the two sets disappears.

A candidate new motif
Finally, the word ATAAGGG/CCCTTAT is a candidate new binding site, since it does not have good comparison scores with any of the known motifs. It scores best with the AFT1 motif, with a 0.52 score which is practically meaningless since 84.9% of all independent 7-letter words score the same or better with at least one motif. It is associated with 13 genes, as shown in Tab. 8, which are overexpressed at late timepoints. The average expression levels for the set and the significance index are shown as a function of time in Figs. 3 and 4.

Comparison with the results of Ref. [5]
As stated in the introduction, the method proposed in Ref.[5] also allows one to identify regulatory motifs without any previous clustering of gene expression data: a linear dependence of the logarithm of the expression levels on the number of repetitions of each regulatory motifs is postulated, and motifs are ranked according to the reduction in $\chi^2$ obtained when such dependence is subtracted from the experimental expression levels. Iteration of the procedure produces a model, that is a set of relevant regulatory motifs, for each expression data set.

We can conclude that the two methods tend to find motifs with a different effect on gene expression: probably the best results can be obtained by using them both on the same data set.

In Ref. [5] such a model is presented for the 14 min. time point in the $\alpha$-synchronized cell-cycle experiment of Spellmann et al., Ref. [17]. We used our algorithm on the same data set to compare the findings. Let us concentrate on the 7-letter words (the longest considered in [5]). We found 9 significant words, reported in Tab. 9. Of these, five coincide with or are very similar to words found by the authors of Ref.[5] (see their Tab. 2). The remaining four (AGGCTAA, GGCTAAG, GCTAAGC and CTAAGCG, whose similarity clearly suggests the existence of a longer motif) are of particular interest for the purpose of comparing the two methods: If one looks at the dependence of the expression levels on the number of occurrences of these words in the 500 bp upstream region, one clearly sees the existence of an activation threshold (see Fig. 5, where such dependence is shown for GCC-TAAG). On the
other hand, by looking at these data one hardly expects a significant reduction in $\chi^2$ when trying to describe this dependence with a straight line. This should be compared to the same dependence for the word AAAATTT, shown in Fig. 6, which is found by both algorithms. On the other hand, there are two 7-word motifs found in [5] that do not pass our significativity threshold, that is CCTCGAC and TAAACAA.

### Table 8: The ORFs in the set S(ATAAGGG) with their expression profiles.

| ORF        | gene | timepoints |
|------------|------|------------|
|            |      | 1  | 2  | 3  | 4  | 5  | 6  | 7  |
| YBR072W    | HSP26| -0.01 | 0.40 | 0.36 | 1.00 | 1.43 | 3.47 | 2.84 |
| YDL133W    |      | -0.04 | 0.32 | -0.34 | -0.25 | -0.56 | -0.22 | -0.32 |
| YDL204W    |      | -0.36 | 0.92 | -0.51 | 0.26 | 0.08 | 4.05 | 3.06 |
| YL136W     | OM45 | -0.97 | -0.27 | 0.21 | -0.25 | 1.32 | 3.47 | 1.79 |
| YLR163C    | MAS1 | 0.04 | -0.01 | 0.11 | -0.01 | 0.08 | 0.30 | -0.03 |
| YLR164W    |      | -0.30 | N/A | -0.27 | 0.06 | -0.18 | 2.19 | 1.69 |
| YLR453C    | RIF2 | -0.07 | -0.27 | 0.32 | -0.01 | 0.71 | 0.69 | 0.08 |
| YML127W    | RSC9 | 0.01 | 0.14 | 0.08 | -0.18 | -0.27 | -0.30 | -1.06 |
| YML128C    |      | -0.12 | 0.20 | 0.97 | 1.56 | 1.36 | 4.32 | 3.47 |
| YNL117W    | MLS1 | -0.30 | -0.04 | 0.71 | -0.30 | -0.27 | 0.76 | 3.18 |
| YPR025C    | CCLI | -0.18 | -0.36 | -0.30 | -0.25 | -0.42 | 0.36 | 0.20 |
| YPR026W    | ATH1 | -0.06 | -0.04 | 0.11 | 0.20 | 0.20 | 0.60 | 1.56 |
| YPR172W    |      | 0.29 | 0.03 | -0.07 | -0.27 | -0.20 | 1.43 | 0.92 |

|            | set average | genome average | significance |
|------------|-------------|----------------|--------------|
| expression (log $2$) | -0.159 | 0.085 | 0.106 | 0.120 | 0.143 | 1.625 | 1.337 |

### Table 9: Significant 7-letter words for the 14-minute timepoint in the $\alpha$-synchronized cell-cycle experiment

| word       | genes | sig |
|------------|-------|-----|
| AAAATTT    | 50    | -7.63 |
| ACGCGTC    | 28    | 6.46 |
| AGATGAG    | 33    | -6.96 |
| GATGAGA    | 25    | -6.47 |
| GAGATGA    | 41    | -6.60 |
| GGCTAAG    | 17    | 7.30 |
| AGGCTAA    | 22    | 6.65 |
| CTAAGGC    | 16    | 6.89 |
| GCTAAGC    | 17    | 6.77 |

**Figure 5**

**expression as a function of occurrences of the word GGCTAAG**
The average expression of genes presenting $n$ occurrences of the word GGCTAAG as a function of $n$ in the 14 min. time point of the $\alpha$-synchronized cell-cycle experiment of Spellmann et al., Ref. [17]. In parentheses is the number of genes with $n$ occurrences of GGCTAAG in the upstream region. The horizontal line represents the average expression for the whole genome.

### Conclusions

We have presented a new computational method to identify regulatory motifs in eukaryotes, suitable to identify those motifs that are effective when repeated many times in the upstream sequence of a gene. The main feature that differentiates our method from existing algorithms for
motif discovery is the fact that genes are grouped a priori based on similarities in their upstream sequences.

Most of the significant words the algorithm finds can be associated to five known regulatory motifs: This fact constitutes a strong validation of the method. Three of them (STRE, MIG1 and UME6) were previously known to be implicated in glucose suppression, while the fact that PAC and RRPE sites are relevant to regulation during the diauxic shift is in agreement with expression coherence data as reported in the web supplement to Ref. [6]. One of the significant words we find (ATAAGGG) cannot be identified with any known motif, and is a candidate new binding site.

It is easy, at least in principle, to extend the method to a larger class of regulatory sites. According to our knowledge of gene regulation, this should be done at least in two directions: (1) the analysis should not be restricted to fixed sequences, but extended to motifs with controlled variability; in particular the extension to spaced dyads [18] should be straightforward; (2) the combinatorial analysis of binding sites [6] could also be performed along the same lines, that is first grouping genes according to which combinations of motifs appear in their upstream region, and then analysing expression profiles within each group.

1Our set is smaller that the one reported in Ref. [2] because we do not allow the upstream sequence to overlap with the previous gene: this eliminates 7 genes from the set.

Acknowledgement

Upstream sequences were downloaded, and many cross-checks were performed, using the impressive collection of packages available from the Regulatory Sequence Analysis Tools (RSAT) [19] at [http://www.ucmb.ulb.ac.be/bioinformatics/rsa-tools/] or ... [http://emnet.cifn.unam.mx/rsa-tools/].