Analysis of Yeast's ORF Upstream Regions by Parallel Processing, Microarrays, and Computational Methods

Steven Hampson
Dept. of Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425
(949) 824-2111
(949) 824-4056 FAX
hampson@ics.uci.edu

Pierre Baldi*
Dept. of Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425
(949) 824-5809
(949) 824-4056 FAX
pfbaldi@ics.uci.edu

Dennis Kibler
Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425
(949) 824-5951
(949) 824-4056 FAX
kibler@ics.uci.edu

Suzanne B. Sandmeyer
Department of Biological Chemistry
College of Medicine
University of California, Irvine
Irvine, CA 92697-1700
(949) 824-7571
sbsandme@uci.edu

Keywords: Gene expression, gene upstream regions, gene regulation, DNA-microarrays, yeast, motifs, Markov models

Abstract

We use a network of workstations to compute all pairwise alignments of the 500 bp upstream regions of 6,225 yeast ORFs (Open Reading Frames). We correlate the alignments with DNA microarray expression data from budding yeast cells over an oxidative stress time course. We confirm on a genomic scale that, in general, genes with extremely similar upstream regions have similar activity levels, even when located on different chromosomes. As the difference in upstream regions increases, the correlation rapidly drops towards zero. Divergent ORFs with overlapping upstream regions do not seem to be correlated in any way. The pairwise alignments coupled with the expression data, together with other computational techniques, suggest a few new putative regulatory binding sites that can be tested experimentally. Finally, we investigate the inherent symmetry present in the two strands of the yeast genome. We show that it extends at least all the way up to 9-mers and is likely to result from different evolutionary pressures operating at different length scales.

*and Department of Biological Chemistry, College of Medicine, University of California, Irvine. To whom all correspondence should be addressed.

1 Introduction

One of the most fundamental questions in biology is how the expression level of tens of thousand of genes is regulated at all times in the life of a cell. A general assumption of gene transcription regulation is that much, although not all, regulation is controlled by the immediately preceding upstream region and is mediated by a complex array of DNA-binding proteins and their cofactors. This has been shown experimentally to a certain extent for a small number of specific genes [Kornberg & Lorch, 1999]. It is also true that the majority of known regulatory motifs are found in the upstream regions. In [van Helden *et al.*, 1998], 99% of the 308 yeast regulatory sites present in the TRANSFAC data base [Wingender *et al.*, 2000] were found to lie within the 800 bp upstream region. To a first order approximation, this general assumption would imply that if two ORFs have very similar upstream regions, then they ought to have similar levels of expression. On the other hand it is also known that in some cases a single base pair change can disrupt and inactivate a regulatory motif, which in turn can have a large impact on gene expression. The availability of complete genomic sequences combined with modern DNA-microarray technology [DeRisi *et al.*, 1999, Holstege *et al.*, 1998, Spellman *et al.*, 1998], probabilistic modeling, and large computing power allows us today to tease out these and other related questions on a full genomic scale in a quantitative way [Zhang, 1999].

Here we study the sequence of 6,225 ORFs in *Saccharomyces cerevisiae* [Goffeau *et al.*, 1996] and their 500 bp upstream regions. In particular, we produce a pairwise alignment and score for all possible pairs of upstream regions. We also experimentally derive expression levels for all yeast ORFs under oxidative stress conditions using DNA microarray gene expression technology. By correlating the alignments with the expression levels we can address the question of whether similarity in the immediately preceding upstream regions does in fact determine similarity in expression levels on a genome wide scale.

A byproduct of this approach identifies a number of DNA protein binding site candidates by several different algorithms. In one approach, we identify a number of possible motif instances by sorting N-mers by their correlation with up or down regulation. The same approach can be used to generalize specific N-mers into more general IUPAC motif descriptions. In a second approach, we look at differential alignments and search for contiguous short stretches of DNA that differ in two upstream regions that are otherwise extremely similar in sequence and very dissimilar in the level of expression of the corresponding genes. A third complementary approach we pursue is the search for short sequences that are not found in *any* upstream region across the entire yeast genome. One possible explanation for the complete absence of a given N-mer in a genome is that, in fact, it binds too efficiently to a given protein.

Finally, it is remarkable that to a first order approximation the nucleotide composition of the yeast upstream regions is symmetric across both strands and given by $P_A = P_T = 30\%$ and $P_C = P_G = 20\%$. We investigate this fact and some of the possible underlying causes by building Markov models of both strands of order up to 8 for both coding and upstream regions.

2 Methods

2.1 Parallel Processing

We used a cluster of 20 workstations running a local distributed message passing algorithm [Kuang *et al.*, 1999] to compute all possible pairwise global alignments of the 6,225 yeast ORF upstream regions. This corresponds to 19,372,200 pairwise alignments. This operation was repeated by taking 100 bp, 200 bp, 300 bp, 400 bp, and 500 bp in the upstream region. As a typical affine gap scoring function, we used +1 for a match, -1 for a mismatch, -2 for initiating a gap, and -1 for prolonging a gap. Thus, for instance, if two 500 bp upstream regions are identical the corresponding score is +500. More generally, if two sequences of length N are identical their score is N. A low score does not necessarily imply that two ORFs are unrelated though, since an exact match over a region of 250 bp followed by a 250 bp complete mismatch region would result in a score of 0.

2.2 Microarrays

Gene expression data was derived by studying the oxidative stress response in yeast using Affimetryx Gene Chip microarray technology [Wodicka, 1997]. In a typical experiment, we used the wildtype yeast strain YPH500 [Sikorski & Hieter, 1989] with 3 untreated controls grown at room temperature and 2 treated data sets, assayed independently. Oxidative stress treatment was given in the form of 0.4mM of oxygen peroxide (H_2O_2) for 5, 10, and 20 minutes. We used GeneChip Expression Analysis v. 3.1 software to obtain the average difference values. All experiments were prepared using the polyA mRNA protocol. Additional experiments and details can be found in [Long *et al.*, 2000].

3 Results

3.1 Alignment Results

In Figure 1 the histogram of the pairwise alignment scores for all the 500 bp upstream regions is given on a logarithmic scale. The lowest score is -184, with one occurrence, the highest is 500 with 41 occurrences, and -44 has the maximum number of occurrences with 580,903. The mean is -44 and the standard deviation is 14. Surprisingly, the same analysis applied to randomly generated data using the first order composition of the upstream regions yields scores that are somewhat higher (average around -33) and therefore less dissimilar on average.

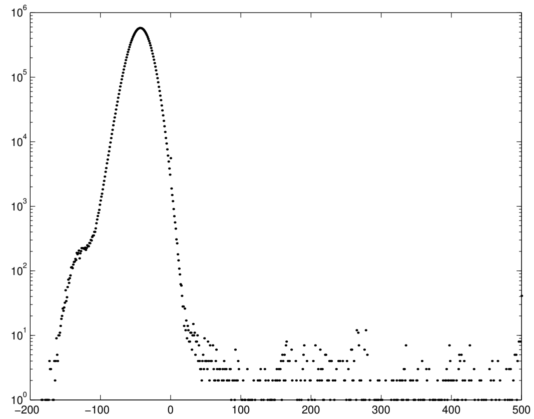

Figure 1: Logarithmic histogram of pairwise alignment scores of all 500 bp upstream regions.

Not counting ORF pairs with overlapping upstream regions, there are 723 pairs with an alignment score of 50 or more, which indicates some degree of homology.

The ORFs in these pairs are not at all unique, however, since some homologous ORFs form families of size 10 or more. These large families are rare, but each could produce $45 = \binom{10}{2}$ or more homologous pairs if all ORFs in the family were mutually homologous. Thus the 732 pairs consist of only 276 ORFs. Based on the first 100 bp only, a score greater than 25 indicates some degree of homology, giving 621 pairs consisting of 214 ORFs, essentially the same ones that showed homology over 500 bp.

3.2 Relation Betweeen Sequence and Expression

Here we study the relationship between alignment scores and expression levels. Expression levels were compared using a number of similarity measures, including Euclidean distance and correlations, computed at individual time steps or across the entire data sets. Because only a relatively small number of genes has significant variations in expression over the treatment period (roughly 10 %), results for individual time steps were essentially the same as for the total set.

The main results are summarized in Table 1. Each entry in the table consists of two numbers. The first component is the number of pairs of upstream regions with an alignment score equal to or greater than a given percentage of the perfect alignment score. The second number is the correlation coefficient computed for all the pairs across all time steps. This is repeated for upstream regions of length 100, 200, 300, 400, and 500. Pairs with overlapping upstream regions or ambiguous expression levels were removed. The entries are "binned" so the 80% line contains only those pairs greater than or equal to 80% and less than 90%, and likewise for the 90% line.

Thus, for instance, there are 24 500 bp-long upstream regions that are identical with an alignment score of exactly 500 and a correlation of 0.996. These 24 pairs, shown in Table 2, correspond to 42 distinct ORFs found on both same and different chromosomes. Similarly, there are 47 500 bp-long pairs with an alignment score between 90% and 99% of the maximum. These exhibit a correlation of 0.327 only. For comparison, a correlation computed using 50 randomly selected pairs of ORFs gives a value of 0.095, close to complete independence.

The relatively large number of identical regions is not too surprising if one takes into account the fact that the yeast genome [Wolfe & Shields, 1997], as well as individual genes may have duplicated during evolution. Inspection of the 24 identical pairs suggests that they occur in several contexts. The list does not include Ty elements, because the LTR sequence upstream of the ATG is less than 500. In addition, delta elements are quite variant and for that reason are not necessarily highly represented in the sets of shorter identical flanking sequence pairs. Only two of the pairs were derived from the blocks of three or greater genes described by Wolfe and Shields. The HMLalpha and MAT alpha loci were among the set. Three sequences represented local expansions containing from two to four copies. In most cases, both the upstream region and coding sequence are identical. However, in one pair, the upstream sequence is identical with a breakpoint for the duplication occurring within the coding region. The majority of the pairs are in telomeric regions of different chromosomes. Two groups of three identical upstream regions account for the fact that there are only 42 different ORFs in the list.

Table 2: The 24 pairs of non-overlapping ORFs with identical 500 bp upstream region.

YAR060C and YHR212C	YAR062W and YHR213W
YBL109W and YHR217C	YBR302C and YML132W
YCL065W and YCR041W	YCL066W and YCR040W
YCL067C and YCR039C	YDL244W and YJR156C
YDL245C and YJR158W	YDL246C and YJR159W
YDR038C and YDR039C	YDR545W and YLR467W
YGR296W and YNL339C	YGR296W and YPL283C
YHR053C and YHR055C	YIL173W and YJL222W
YIL176C and YJL223C	YLR155C and YLR158C
YLR155C and YLR160C	YLR158C and YLR160C
YLR159W and YLR161W	YNL339C and YPL283C
YOR393W and YPL281C	YOR394W and YPL282C

The table confirms that very similar upstream regions are associated with similar expression levels, although the correlation falls off quite rapidly as the similarity decreases. Similar results are seen at each individual time step as well as with different similarity measures. Thus, besides confirming that highly similar upstream regions produce similar expression levels, it can be concluded that homologous ORFs, presumably resulting from ORF duplication, do not need to diverge greatly before producing very different expression patterns. This is somewhat surprising in that homologous ORFs, which likely produce similar proteins, might be expected to be correlated in both function and expression.

Identical 100 bp upstream regions ensure a very high degree of correlation, and small variations in this region reduce the correlation essentially to 0. However, it is not correct to infer from this that most of the regulation is concentrated in the first 100 bp because, in general, sequences that have identical 100 bp regions tend to have identical or similar upstream regions at longer lengths. Strings that are similar in the first x hundred bp are very often similar to some extent in the next 100 bp, thus it is impossible to localize such effects any further. By a similar argument, ORFs that are similar in their

Table 1: Binned number of sequence pairs and correlation in expression, for different upstream region lengths and different sequence similarity cutoffs (cf. main text).

	100	200	300	400	500
100%	(81,0.939)	(45,0.980)	(36,0.981)	(25,0.996)	(24,0.996)
90%	(83,-0.132)	(71,0.216)	(62,0.517)	(54,0.079)	(47,0.327)
80%	(47,0.026)	(62,0.100)	(62,0.028)	(43,0.184)	(47,0.219)

500 bp upstream regions are likely to be similar over an even larger context. We did look at ORFs with different 500 bp upstream region that were identical on the first 100 bp but they were too rare to derive any statistically significant conclusions.

We also tried to look at the correlations with respect to time, i.e. trying to detect correlations between temporal patterns of expression. Unfortunately such a correlation cannot be detected with reliable statistics here because there are only about 600 ORFs that exhibit significant changes in the oxidative stress experiment, and the number of highly similar upstream regions is already small. In particular, two genes that fluctuate more or less randomly near the background level are highly correlated by the measure above, in that they are both essentially unexpressed, but may very well have completely uncorrelated activities as a function of time.

Because some transcription factors operate symmetrically on both strands, one may be tempted to hypothesize that some positive correlation may exist between the level of expression of genes located on different strands but with large overlapping upstream regions. On the other hand, one could also hypothesize that expression of one member of the pair physically inhibits expression of the other one and therefore they ought to be negatively correlated.

We found 1,350 divergent ORF pairs with some degree of overlap in their 500 bp upstream regions, that is with their ORF origins within 1,000 of each other and with an orientation that produces overlap of upstream regions. Of these, 44 pairs have their ORF origins between 490 and 510 of each other, resulting in an overlap in their 500 bp upstream region of at least 490. The correlation of these diverging ORFs with highly overlapping upstream regions is -0.08, close to random. More generally, the number of divergent ORF pairs and their expression correlation are given in Table 3, for all possible distances between the two ORF origins between 0 and 1000, in blocks of 100. There is no evidence of correlation, positive or negative, in expression. The 0.52 correlation at the 700-800 distance is spurious and due to a single outlier associated with the divergent pair YOL039W and YOL040C which have extremely high levels of expression (in the 10-20,000 range, almost two orders of magnitude above the average). We also classfied ORF's into three classes–decrease/no change/increase–and again did not find any correlation for any level of overlap.

Thus, in general, when two ORFs have almost identical upstream regions their expression levels are correlated, even when they are on different chromosomes. But if two neighboring ORFs are divergent, their expression levels do not seem to be correlated, even when they have highly overlapping 500 bp upstream regions.

Table 3: Number of divergent ORF pairs and correlation between the expression level of each pair, as a function of the distance in base pairs between their origins.

distance	number	correlation
0-100	33	-0.13
100-200	100	-0.03
200-300	286	-0.03
300-400	172	-0.04
400-500	152	-0.00
500-600	143	0.01
600-700	155	0.10
700-800	124	0.52
800-900	105	0.04
900-1000	80	-0.04

3.3 Expression and Regulatory Motifs

We have explored a number of techniques for identifying regulatory motifs in the upstream region [Hu *et al.*, 2000] and report here the results of one of these approaches based on probabilistic over-representation [van Helden *et al.*, 1998]. In this approach, possible motif instances are identified by looking for strings that are over-represented in an experimentally derived test set of ORFs which respond in a similar fashion to a shift in growth conditions. The expectation is that over-represented strings in the set are apt to be causally related to the observed pattern of expression that defines the set. Specifically, the observed number of occurrences of a string in all upstream regions is used to calculate the probability of seeing X or more occurrences in the test set by chance, where X is the observed count in the test set. The smaller the probability, the greater the degree of over-representation.

We have written a program that can rapidly search the space of all possible N-mer strings using binary encoding of each nucleotide and associating each N-mer with a number that is also the index of an array. For $N < 10$, the yeast genome can be exhaustively analysed in less than a minute, so finding the most over-represented strings is guaranteed.

In this analysis, an ORF was classified as significantly changed if the average of its 5 and 10 min expression levels was more than 1.5 fold different from its time 0 expression level. A background level of 20 was added to each expression value before the fold change was calculated in order to eliminate incorrect classification based on random background differences in essentially unexpressed ORFs. Using this method, 1,556 ORFs were classified as changed, 736 up and 820 down. The remaining ORFs were removed from the analysis. Other classification rules give different numbers in each set, and the unchanged ORFs can be included in the analysis, but the final motif-finding results are quite similar over a range of variations in the approach. In fact, a statistical analysis using the tools in [Long *et al.*, 2000] suggests that only about 600 ORFs are significantly changed, and of those about 200 are expected to be false positives, so the results reported here provide some evidence that the approach works despite the presence of a considerable amount of noise.

Two approaches were taken using these data. In the first case, the UP set was used as the test set and UP+DOWN as the comparison set. Alternatively, the set of all ORFs (UP+DOWN+UNCHANGED) can be used as the comparison set. This corresponds most closely to the approach used in [van Helden *et al.*, 1998]. In the second approach, rather than asking if a string is over-represented in the UP set, the set of ORFs containing each N-mer string was used as the test set. The over-representation of UP ORFs in that set was then calculated. If a string's presence is correlated with up regulation, both statistics should be positive, and they do in fact give similar results. Results from the latter approach are reported here.

Evidence that the approach works is provided by the fact that for 5-mers, the first and third best strings identified by this method are the well-known stress element CCCCT and its reverse complement AGGGG [Martinez-Pastor *et al.*, 1996]. Other 5-mers on the list may be effective in isolation like the stress element, but they are more apt to be portions of larger patterns (Table 4, left column).

Under the assumption that regulatory motifs should be effective in either orientation, the strings can be sorted by over-representation for combinations of each string and its reverse complement. In this case the stress element is first on the list (Table 4, right column).

Table 4: Ranked list of top 20 5-mers for up-regulation, first at the top (left column). The stress element CCCCT comes first. Its reverse complement comes third. Ranked list of top 10 5-mers for up regulation when combined with their reverse complement (right column).

CCCCT	**AGGGG**
ACCCC	**CCCCT**
AGGGG	ACCCC
CACCC	GGGGT
AAGGG	AAGGG
CCACA	CCCTT
GCCCC	GGGTG
GGGGA	CACCC
GGGAG	TCCCC
CTGGA	GGGGA
AACCC	GGGAG
CCCCC	CTCCC
GGGGT	GGGGC
GGGGG	GCCCC
CCCTT	GGGGG
TCTCC	CCCCC
CTCCC	CGGGG
TCCCC	CCCCG
CCCCG	TGTGG
AGGGA	CCACA

Further evidence for the effectiveness of this method is provided by the fact that for 7-mers the YAP1 element (TTACTAA) and its reverse complement (TTAGTAA) are first and seventh on the list (Table 5, left column). In yeast the bZIP transcription factor YAP1 has been implicated in the response to oxidative stress. This protein preferentially binds the sequence TTACTAA, which is found in the upstream flanking region of a number of genes whose transcription increases in response to stress induced by oxidative species such as hydrogen peroxide [Wu & Moye-Rowley, 1994, Fernandes *et al.*, 1997, Coleman *et al.*, 1999]. The alternative method of measuring over-representation, using the UP set as the test set, ranks the YAP1 element and its reverse complement as first and second. So while the results are generally similar, one approach may ultimately prove superior to the other. Several strings on the list contain the stress element or its reverse complement, which may explain their correlation with increased expression, but others may reflect novel motifs. Combining each string with its reverse complement, the YAP1 element is first on the list (Table 5, right column).

Strings that are predictive of down regulation are also interesting (Table 6). We do not have a known down

Table 5: Ranked list of top 20 7-mers for up-regulation, first at the top (left column). The YAP element comes first. Its reverse complement comes seventh. Ranked list of top 10 7-mers for up regulation when combined with their reverse complement (right column).

TTACTAA	**TTAGTAA**
ACCCACG	**TTACTAA**
ATTACTA	TTAGGGG
AACTCCG	CCCCTAA
GTGTGTG	TACGTAA
CACCCCT	TTACGTA
TTAGTAA	AACTCCG
AAAAGGG	CGGAGTT
TGCCTAT	GTGTGTG
CACAAAC	CACACAC
GAGTGTG	AAAGGGG
CCGTGGA	CCCCTTT
GGCAGGT	AAGGGGT
ACCCCTT	ACCCCTT
CCCACAC	TGTGTGG
CTGCCTG	CCACACA
TTAGGGG	GGGGCTG
AGGTTAC	CAGCCCC
AGGGAAC	AAAAGGG
CTTCCGT	CCCTTTT

regulator that should appear on the list, but those that do are quite compelling in that, generally, both the string and its reverse complement are strongly associated with down regulation. In addition, both are generally localized in the 100 to 200 bp upstream region, most likely corresponding to the preferred region of localization observed in [Hughes *et al.*, 2000]. For example, a histogram of the location of the first string AAAATTT is shown in Figure 2. Most strings on the list show similar localization. Further analysis of these strings indicates that most are portions of a small number of longer, degenerate motifs.

The main advantage of this approach is that the statistical properties of individual N-mers can be quickly and exhaustively computed, thus rapidly identifying strings of interest. The main drawback is that in some cases a single degenerate motif can result in a large number of motif instances, a fact that may not be apparent from the intermixed list of all over-represented N-mers. Identifying one or more of these motif instances is thus only a first step in the characterization of a complete motif. In some cases it is relatively easy to assemble a complete motif by hand [van Helden *et al.*, 1998], but we are developing automatic tools for this process.

For example, the general approach of computing over-representation does not have to be applied only to individual strings. This was shown in the previous example of computing a value for pairs of strings; the string itself plus its reverse complement. This made sense since combining the string and its reverse complement is the same as counting the string itself on both strands of the DNA. In fact the approach can be further generalized to any set of strings, such as those resulting from candidate motif descriptions. We have explored this possibility us-

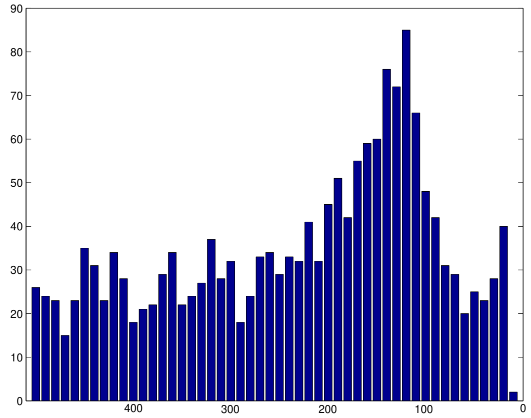

Figure 2: Localization of AAAATTT in the 100-200 bp upstream region.

Table 6: Ranked list of top 20 7-mers for down-regulation, first at the top (left column). Ranked list of top 10 7-mers for down regulation when combined with their reverse complement (right column).

AAAATTT	AAAATTT
AAATTTT	AAATTTT
AATTTTT	AAAAATT
CGATGAG	AATTTTT
AAAAATT	CTCATCG
GCGATGA	CGATGAG
GAAAAAT	TCTCATC
GATGAGC	GATGAGA
TGAAAAA	TCATCGC
GAGATGA	GCGATGA
TCTCATC	TCATCTC
ATGAGCT	GAGATGA
CTCATCG	GATGAGC
GATGAGA	GCTCATC
CATAGTT	ATTTTTC
TTTTTTT	GAAAAAT
AAAGAAA	TTTTTCA
AGATGAG	TGAAAAA
AACGCGA	ATGAGCT
TTTTATC	AGCTCAT

ing both IUPAC and weight matrix motif representation. The value of this is that if multiple motif instances with equal over-representation values can be combined, the over-representation of the resulting set is greater than the over-representation of the individual instances that make it up. This makes it possible to identify an over-represented motif, even if none of its individual motif instances are highly over-represented.

Another useful set of instances for which over-representation can be computed is the $m1$ (miss = 1) "shell" around each $m0$ (exact match, miss = 0) N-mer. That is, while the over-representation measure is normally calculated for each possible $m0$ string, it can also be calculated for the $m1$ set of strings that surround each $m0$ string and differ in exactly one position. For a length N string, there are $3N$ strings in this shell. The expectation is that because of possible variations in a motif, if an over-represented motif instance is found, its $m1$ shell is apt to be over-represented as well. This is especially true for the most central or "prototypic" instances of the motif. For motifs with a large amount of variability, the $m1$ shell can be more over-represented than the $m0$ center or any of the individual strings that make up the $m1$ shell. This is especially important with longer motifs ($N > 6$). Over-representation of the $m1$ shell provides an independent measure of the importance of the $m0$ instance. If both $m0$ and $m1$ are highly over-represented, the chances of $m0$ being a real motif instance are increased. Alternatively, the $m0$ and $m1$ strings can be combined and a single over-representation score computed. Whether this is useful or not depends on the degree of variability of the motif. If the motif has low variability, the over-representation of the $m1$ shell will be correspondingly low. The over-representation of the $m2$ shell (exactly two mismatches) can also be calculated, but empirically it was not particularly useful.

This approach can also be applied to find more selective generalizations of a single motif instance. Specifically, an IUPAC motif description permits an arbitrary disjunction of bases at each position of the motif. So, given a single motif instance, all possible generalizations can be tried at each position to see which resulting set of strings has the highest over-representation value. This can be applied sequentially and repeatedly to each position in the motif until no further improvement is possible. For example, if 9-mers are analyzed, the string GCGATGAGC is identified as a possible down regulator since it is contained in 26 DOWN ORFs and only 3 UP. The corresponding counts for its reverse complement are 20 and 2, supporting the hypothesis that it is a regulatory motif instance. Both the string and its reverse complement are highly localized in the 100-200 bp region, and both have over-represented $m1$ shells. Applying the IUPAC hill-climbing algorithm to this string results in the more general motif [TGC][AC]GATGAG[AGC] with 222 down and 48 up. Many of the 18 9-mers, plus their reverse complements, generated by this motif are individually identifiable as candidate motif instances, but the IUPAC hill-climbing algorithm eliminates the necessity of grouping and generalized the appropriate instances by hand. Analysis of overlapping 9-mers suggests that the sequence of conserved bases is somewhat longer and, as a minimum, should include an additional T at the right end. In the context of oxidative stress, this appears to be a powerful inhibitory motif [Luche *et al.*, 1993].

3.4 Differential Alignments and Binding Factors

When genes have almost identical upstream regions but different levels of expression, it is reasonable to suspect that at least in some cases the difference results from the few bases that are different in the two upstream sequences. In addition, it is of interest to check if the alignment of the two upstream sequences reveals that the sequence differences are concentrated in a short contiguous regions, rather than being spread out over the 500 bp. When this is the case, one can further hypothesize that the short contiguous regions may correspond to a protein binding site and it is the interaction with such proteins that result in the difference in gene expression. Notice that this is plausible but not necessary. Furthermore, the contiguous region may not be perfectly centered on a protein binding site, but only display some degree of overlap with it, since the disruption of a third or half of a binding sequence is also likely to disrupt protein interactions. Thus it is not possible to associate protein binding site boundaries with gap alignment boundaries in any precise form. Any mismatch, even one or a few bp long, between two almost identical sequences that have very different expression levels signals a location where regulatory proteins may interact.

Examples of short sequences isolated by this method include CTTATTAAC and TGCTTGAAGG. Both are not found in the TRANSFAC data base. A more systematic analysis of such sequences, and their immediate context, is in progress. In particular, the analysis can be further complemented by looking at the DNA structural properties of the corresponding sequences using the techniques described in [Baldi *et al.*, 1999].

3.5 Missing N-Mers

It is also of interest to study all the N-mers that are completely absent from the entire yeast genome, or from the upstream regions only. For low but non-trivial values of N, such regions ought to be very rare and one may conjecture that selection may have operated against them. This could occur for a number of reasons. One is that the corresponding DNA has structural properties (e.g. cruciform) that may interfere with basic cellular processes, susch as DNA transcription or replication. Another possibility, is that such regions may in fact correspond to protein binding factors that are *too good*, in the sense that the dissociation constant is too low so that once bound the corresponding protein never comes off. It is clear that once such sequences are found, their putative negative effects on the cell can be tested experimentally by inserting them into genomic DNA.

Using the program described above, we find that in yeast there are no missing N- mers for N up to 8, across the entire genome. If we look at all the 500 bp upstream regions only, then there is a single missing 8-mer: CCCGAGCC on the direct strand. It is however found 8 times in the reverse complement upstream regions. There are 16 8-mers, all very CG rich, that are found a single time across all the 500 bp upstream regions. Of these, two stand out because they are found also a single time across the reverse complement upstream regions. These are: TCCCCGGG and CCCCGGGG.

Structural measures of these 8-mer computed using a number of structural scales as in [Baldi *et al.*, 1999, Baldi & Baisnée, 2000] are given in Table 7 in standard deviation units, showing that at least in some cases these 8-mer may have fairly unusual structural properties.

We can rapidly look at the statistical significance of these results. Using the first order composition $P_C \approx P_G \approx 0.2$ alone, and assuming on the order of $6,000 \times 500$ well defined upstream regions, the expected number of occurrences of each 8-mer is roughly 7. Thus, the fact that we find one missing 8-mer is not surprising. What is actually surprising, is that we do not find more than one. Additional tests are in progress, including on larger genomes where statistics may be even more significant.

Table 7: Structural measures of rare 8-mers (see text) in standard deviation units. BS = Base Stacking Energy, PT=Propeller Twist Angle, PD = Protein Deformability, B= Bendability, PP=Position Preference.

8-mer	BS	PT	PD	B	PP
CCCGAGCC	-1.286	1.567	0.978	0.373	1.479
TCCCCGGG	-0.600	2.477	1.931	-0.598	-1.242
CCCCGGGG	-0.381	3.028	2.181	-0.679	-0.955

3.6 Yeast Upstream Region Symmetries

We have seen that the overall first order composition of the yeast ORF upstream regions is symmetric or, in other words, both the direct and reverse complement strand have the same first order Markov model. In particular, the exact parameters of the first order Markov model (i.e. average composition) for these upstream regions are given in Table 8. The symmetry is obvious in the fact that $P_A \approx P_T$ and $P_C \approx P_G$.

Table 8: First order distribution of yeast 500 bp upstream regions.

A	0.3166
T	0.3128
G	0.1834
C	0.1872

This symmetry is notable and is found in other organisms, such as *Escherichia coli* which has an even more symmetric first order distribution over its entire genome $P_A = 0.2462$, $P_T = 0.2459$, $P_C = 0.2542$ and $P_G = .2537$. This symmetry extends to higher order statistics. But what it is even more remarkable, is that once the first order symmetry is factored out, a high degree of symmetry is still present. This can be shown in a number of ways. All the results in this section are computed after discarding the 1,350 divergent ORFs with overlapping 500 bp upstream regions, since keeping each pair would introduce additional unrepresentative symmetries.

First, we constructed Markov models (for instance [Baldi & Brunak, 1998]) of all orders up to 9 for the 500 bp upstream regions. A Markov model of order N has 4^N parameters associated with the transition probabilities $P(X_N|X_1,\ldots,X_{N-1})$ also denoted $(P(X_1,\ldots,X_{N-1}) \rightarrow X_N))$ for all possible $X_1,\ldots,X_N$ in the alphabet together with a starting distribution of the form $\pi(X_1,\ldots,X_{N-1})$. In our case, beyond $N = 9$, the number of parameters in the model becomes too large with respect to the number of data points (roughly $500 \times 5000 = 2,500,000$) to ensure a proper fit of the model to the data.

In this context, we say that a Markov model of order N is symmetric if it is identical to the Markov model of order N of the reverse complement. Because of the complementarity between the strands, a Markov model of order N of one strands immediately defines a Markov model of order N on the reverse complement. Thus a Markov model is symmetric if and only if $P(X_1 \ldots X_N) = P(\bar{X}_N \ldots \bar{X}_1)$. A Markov model of order N induces an equilibrium distribution over lower order N-mers. In particular, a Markov model of order N induces a first order equilibrium distribution that must satisfy the balance equation:

$$P(X_2,\ldots,X_N) \quad = \quad \sum_Y P(X_N|Y,X_2\ldots,X_{N-1})$$

$$P(Y, X_2, \ldots, X_{N-1}) \quad (1)$$

If a Markov model of order N is symmetric, its restrictions or projections to lower orders are also symmetric. The converse, however is not true. In general, a symmetric Markov model of order N can have multiple non-necessarily symmetric extensions to a Markov model of order $N+1$ or higher. Thus the fact that the first order distribution of yeast is symmetric does not necessarily imply that the second order distribution is also symmetric. But this is precisely the case.

The parameters of the corresponding Markov model of order 2 are given in Table 9. It is easy to check by inspection that they correspond to models that are symmetric with respect to the reverse complement strand. For instance AG on the direct strand corresponds to CT on the reverse complement and $P(AG) = 0.0589$ while $P(CT) = 0.0583$. Likewise for the third order model, for instance, $P(CAG) = 0.0115 \approx P(CTG) = 0.0116$. While displaying the parameters of the Markov models beyond order 2 would take too much space, we observe such symmetry in orders up to 9.

Table 9: Second order transition parameters and dinucleotide distribution of yeast 500 bp upstream regions.

A → A	0.3643	AA	0.1154
A → T	0.2806	AT	0.0889
A → G	0.1858	AG	0.0589
A → C	0.1684	AC	0.0533
T → A	0.2602	TA	0.0814
T → T	0.3662	TT	0.1146
T → G	0.1858	TG	0.0581
T → C	0.1882	TC	0.0589
G → A	0.3166	GA	0.0581
G → T	0.2784	GT	0.0511
G → G	0.1945	GG	0.0357
G → C	0.2106	GC	0.0387
C → A	0.3304	CA	0.0619
C → T	0.3116	CT	0.0583
C → G	0.1639	CG	0.0307
C → C	0.1941	CC	0.0364

We can then test how well the Markov models fit the upstream data and this is shown in Table 10, displaying the correlation between the expected counts and the actual counts over the set of all N-mers (N = 5, 7, and 9), where the expected counts are produced by Markov models of order $K = 1, \ldots, \leq N$.

To test the difference in behavior between yeast data and artificial random data, we generated a long string using a randomly constructed Markov model of order 5. Markov models of order 1 to 5 were fitted to this random data. Correlations between counts and expected counts of 5-mers, 7-mers, and 9-mers for these models are displayed in Table 11.

Table 10: Correlation of counts (C) to expected counts($E(C)$) produced by Markov models of order $K = 1, \ldots, N$ measured on yeast 5-mers, 7-mers, and 9-mers.

order	5-mers	7-mers	9-mers
1	.82	.72	.57
2	.91	.83	.67
3	.97	.91	.77
4	.99	.95	.83
5	1.00	.98	.88
6		.99	.93
7		1.00	.97
8			.99
9			1.00

Table 11: Correlation of counts (C) to expected counts($E(C)$) produced by Markov models of order $K = 1, \ldots, 5$ to artificial data produced by a random Markov model of order 5.

order	5mers	7mers	9mers
1	.12	.07	.06
2	.17	.11	.08
3	.36	.25	.19
4	.68	.54	.46
5	1.00	.99	.99

A number of significant observations can be made. First, a random Markov model of order 5 perfectly fits the statistics of the 5-mers it generates (for large data sets), which is trivial, but it also fits quite well the statistics of the 7-mers and 9-mers it can generate. In contrast, the yeast Markov model of order 5 does not fit well the 9-mers in yeast. Thus there are long range (> 5) constraints operating on actual DNA. On the other hand, a fourth order Markov model does not fit the random fifth order data very well, since the correlation is 0.68. The yeast data fourth order Markov model, however, exhibits a correlation of 0.99 with yeast fifth order data. Thus the convexity profile of the curves obtained with artificial data is very different from the case of yeast sequences. In particular, low order models perform significantly worst on random data of order 5. All together these results suggest that the yeast data is closer to a mixture of low and high order constraints.

A complementary approach to testing the symmetry at different orders is to consider the correlation between the counts C of N-mers, for each value of N (N = 5, 7, and 9), across both strands. Furthermore, for each N this correlation can be computed after "factoring out" Markov models of order $K < N$. The lower order Markov models can be factored out by computing the corresponding expected counts $E(C)$, and using the ratio $C/E(C)$ or the difference $C - E(C)$ in the computation of the correlations. The purpose of this operation is to analyze whether there is any symmetry left once the effect of lower order constraints is taken into account. Indeed, it is essential to understand that if data is generated randomly using a symmetric Markov model of order 1, it will exhibit higher order symmetries since, for instance, $P(AAA) = P(A)^3 = P(T)^3 = P(TTT)$. However if we factor out the first order Markov model as described above, such higher order symmetry will disappear entirely.

The results are displayed in Table 12 for the ratio and in Table 13 for the difference. The first row of Table 12 displays correlation between the raw counts without dividing by any expectation. This row shows how high the degree of symmetry is. The following rows show what happens when Markov models of increasing order are factored out. Once again symmetry seems to result from constraints operating at different length scales and acting equally on both strands of DNA.

Table 12: Row 0 represents the correlation between the counts C of N-mers ($N = 5, 7, 9$) between the direct upstream strand and its reverse complement. In rows $K = 1$ to 9, similar correlations are computed but using the ratio $C/E(C)$ where $E(C)$ is the *expected* number of counts produced by a Markov model of order K fitted to the upstream regions.

order	5-mers	7-mers	9-mers
0	.99	.99	.95
1	.97	.90	.55
2	.94	.83	.45
3	.94	.77	.36
4	.82	.57	.24
5	*	.45	.18
6		.34	.14
7		*	.10
8			.09
9			*

While the preceding analysis applies to the upstream region as a whole, this region is not entirely homogeneous. In Table 14, we display the average first order composition of the 500 bp upstream region calculated in successive bins of length 50. The distribution of C and G is fairly uniform, hence symmetric, across the upstream region. There is more variability and less symmetry in A

Table 13: Same as previous table. In rows $K = 1$ to 9, similar correlations are computed but using the difference $C - E(C)$ where $E(C)$ is the *expected* number of counts produced by a Markov model of order K fitted to the upstream regions.

order	5-mers	7-mers	9-mers
1	.98	.97	.93
2	.98	.96	.92
3	.97	.95	.90
4	.94	.93	.87
5	*	.88	.83
6		.78	.73
7		*	.46
8			.27
9			*

and T. Not surprisingly, there are more A's in the 0-50 bin due to the well known A-rich signals in the immediate upstream portion of an ORF. A and T are equally frequent in the 100-150 region and, interestingly, the effect is reversed in the 150-200 region where T is relatively high and A somewhat lower. This first order effect is also corroborated by the fifth order effect seen in Table 15, displaying the correlation of counts between the leading strand and the reverse complement. The 0-50 and 150-200 regions stand out as being the least symmetric.

Table 14: First order composition of the upstream regions at different length intervals.

length	A	T	G	C
0-50	0.3823	0.2691	0.1703	0.1783
50-100	0.3466	0.3101	0.1749	0.1683
100-150	0.3217	0.3241	0.1787	0.1755
150-200	0.2966	0.3394	0.1782	0.1858
200-250	0.3010	0.3311	0.1799	0.1881
250-300	0.3045	0.3189	0.1848	0.1919
300-350	0.3025	0.3149	0.1893	0.1933
350-400	0.3013	0.3106	0.1918	0.1963
400-450	0.3047	0.3071	0.1917	0.1964
450-500	0.3052	0.3028	0.1942	0.1979

We have also investigated symmetry in coding regions, as well as on all 16 chromosmes and on the mitochondrial DNA. As expected, coding and upstream regions have different Markov models: a Markov model of order N of the coding regions does not fit well the upstream regions, and furthermore the coding regions are less symmetric. In spite of that, however, we still find strong symmetric constraints operating within coding regions. Overall the 16 chromosomes are all very symmetric and similar

Table 15: Correlation of 5-mers in the leading and reverse complement upstream region in 50-base blocks.

0-50	0.58
50-100	0.92
100-150	0.99
150-200	0.89
200-250	0.94
250-300	0.98
300-350	0.98
350-400	0.98
400-450	0.99
450-500	0.99

in composition. The mitochondrial DNA is also roughly symmetric, at least to a first order, but with a significantly different composition ($P_A = 0.42$, $P_T = 0.41$, $P_C = 0.08$, and $P_G = 0.09$). Further work in this area is in progress.

4 Discussion

In summary, we have first investigated the notion that, for most genes, gene regulation depends predominantly on the upstream sequence on a genome wide scale. Our basic approach consists in correlating all pairwise alignments of upstream regions with expression data. In particular, to a first approximation ORFs with highly similar upstream regions have similar expression levels. Small changes in upstream regions, however, rapidly dilute the correlation to random levels. We find no correlation between divergent ORFs with any degree of overlap in their upstream regions.

We have developed novel techniques for identifying putative protein binding sites and regulatory regions on a genomic scale. More specifically, we have filtered upstream N-mers on the basis of at least three new criteria: (1) correlation to expression levels; (2) differential alignment of upstream regions that are similar in sequence but dissimilar in expression; and (3) under-representation. By these methods, we have found several putative binding sites not found in the TRANSFAC database. The identification of the YAP1 binding site within the upstream flank of genes that are positively regulated under oxidative stress conditions confirms the effectiveness of our computational approach for the identification of candidate regulatory motifs. In addition, this study expands the number of genes known to be regulated by oxidative stress that are potential targets of YAP1 action.

Finally we have investigated the remarkable composition symmetry of the yeast upstream regions as well as entire genome. In particular, we have shown that such symmetries extend considerably beyond the average composition and are likely to result from evolutionary pressures operating over different length scales.

The present work can be extended in several directions which are currently in progress. Homology data for the coding region is available at the protein level [Wolfe & Shields, 1997] and can be computed at the DNA level. In conjunction with the upstream DNA homology data, this provides a rich opportunity for the construction of possible ORF phylogenies and the comparison of divergence rates for upstream and downstream regions. Additional gene expression data, including publicly available data, can be used in the correlation analysis. Additional data obtained over *different* responses, involving a different set of genes, will allow us to look at temporal correlations more closely. Likewise our correlation results can be further strengthened by looking at other model organisms. Some of the putative binding sites ought to be tested experimentally and statistical studies of missing N-mers, as well as higher order genomic symmetries, should be extended to other organisms. It is remarkable to see that technological progress, both on the biological and computational side, now enables us to probe these and many other questions on a genomic scale.

Acknowledgements

We thank the UCI Computational Genomics Group for many helpful discussions. Array experiments were supported by a gift to the Chao Family Comprehensive Cancer Center and by a Howard Hughes Medical Institute Research Resources Grant. We thank L. Yieh and S. Trinidad for assistance with the GeneChip Experiments. L. Yieh was supported on a Biomedical Research and Education Training Grant to SBS. The work of PB was supported by a Laurel Wilkening Faculty Innovation ward.

References

[Baldi & Baisnée, 2000] Baldi, P. & Baisnée, P. F. (2000). Computational analysis of DNA structure for sequences and repeats of all lengths. Submitted.

[Baldi & Brunak, 1998] Baldi, P. & Brunak, S. (1998). Bioinformatics: The Machine Learning Approach. MIT Press, Cambridge, MA.

[Baldi *et al.*, 1999] Baldi, P., Brunak, S., Chauvin, Y. & Pedersen, A. G. (1999). Structural basis for triplet repreat disorders: a computational analysis. *Bioinformatics*, **15**, 918–929.

[Coleman *et al.*, 1999] Coleman, S. T., Epping, E. A., Steggerda, S. M. & Moye-Rowley, W. S. (1999).

Yap1p activates gene transcription in an oxidant-specific fashion. *Mol. Cell. Biol.*, **19**, 8302–8313.

[DeRisi *et al.*, 1999] DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1999). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686. In press.

[Fernandes *et al.*, 1997] Fernandes, L., Rodrigues-Pousada, C. & Struhl, K. (1997). Yap, a novel family of eight bzip proteins in Saccharomyces cerevisiae with distinct biological functions. *Mol. Cell Biol.*, **17**, 6982–6993.

[Goffeau *et al.*, 1996] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Hoheisel, J. D., Jacq, C., Johnston, M., adn H. W. Mewes, E. J. L., Murakakmi, Y., Philippsen, P., Tettelin, H. & Oliver, S. G. (1996). Life with 6000 genes. *Science*, **274**, 546–567.

[Holstege *et al.*, 1998] Holstege, F. C. P., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S. & Young, R. A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.

[Hu *et al.*, 2000] Hu, Y., Sandmeyer, S., Laughlin, C. M. & Kibler, D. (2000). Combinatorial motif analysis and hypothesis generation on a genomic scale. *Bioinformatics*. In press.

[Hughes *et al.*, 2000] Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. (2000). Mcomputational identification of *cis*-regulatory elements associated with groups of functionally related genes in *saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.

[Kornberg & Lorch, 1999] Kornberg, R. D. & Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, **98**, 285–294.

[Kuang *et al.*, 1999] Kuang, H., Bic, L. L., Dillencourt, M. B. & Chang, A. C. (1999). Podc: Paradigm-oriented distributed computing. In *7th IEEE Workshop on Future Trends of Distributed Computing Systems (FTDCS'99), Capetown, South Africa*.

[Long *et al.*, 2000] Long, A. D., Mangalam, H. J. & Baldi, P. (2000). Cyber-t: a set of statistical tools for the analyis of high-density array data. *Genome Research*. Submitted.

[Luche *et al.*, 1993] Luche, R. M., Smart, W. C., Marion, T., Tillman, M., Sumrada, R. A. & Cooper, T. G. (1993). Saccharomyces cerevisiae BUF protein binds to sequences participating in DNA replication in addition to those mediating transcriptional repression (URS1) and activation. *Mol. Cell. Biol.*, **13**, 5749–5761.

[Martinez-Pastor *et al.*, 1996] Martinez-Pastor, M. T., Marchler, G., Schuller, C., Marchler-Bauer, A., Ruis, H. & Estruch, F. (1996). The *saccharomyces cerevisiae* zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress-response element (STRE. *EMBO Journal*, **15**, 2227–2235.

[Sikorski & Hieter, 1989] Sikorski, R. S. & Hieter, P. (1989). A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *saccharomyces cerevisiae*. *Genetics*, **122**, 19–27.

[Spellman *et al.*, 1998] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, **9**, 3273–3297.

[van Helden *et al.*, 1998] van Helden, J., Andre, B. & Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.

[Wingender *et al.*, 2000] Wingender, E., Chen, X., Hehl, R., Karas, H., liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I. & Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.

[Wodicka, 1997] Wodicka, L. H. (1997). Genome-wide expression monitoring in *saccharomyces cerevisiae*. *Nature Biotechnology*, **15**, 1359–1367.

[Wolfe & Shields, 1997] Wolfe, K. H. & Shields, D. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**, 708–713.

[Wu & Moye-Rowley, 1994] Wu, A. L. & Moye-Rowley, W. S. (1994). GSH1 which encodes gamma-glutamylcysteine synthetase is a target gene for YAP-1 transcriptional regulation. *Mol. Cell. Biol.*, **14**, 5832–5839.

[Zhang, 1999] Zhang, M. Q. (1999). Large-scale gene expression data analysis: a new challenge to computational biologists. *Genome Research*, **9**, 681–688.