BIOINFORMATICS

Vol. 18 no. 4 2002
Pages 513–528

Distribution patterns of over-represented k-mers in non-coding yeast DNA

*Steven Hampson, Dennis Kibler and Pierre Baldi**

Department of Information and Computer Science, Institute for Genomics and Bioinformatics, University of California, Irvine, Irvine, CA 92697-3425, USA

Received on May 24, 2001; revised on October 21 and November 30, 2001; accepted on December 5, 2001

ABSTRACT

Motivation: Over-represented k-mers in genomic DNA regions are often of particular biological interest. For example, over-represented k-mers in co-regulated families of genes are associated with the DNA binding sites of transcription factors. To measure over-representation, we introduce a statistical background model based on single-mismatches, and apply it to the pooled 500 bp ORF Upstream Regions (USRs) of yeast. More importantly, we investigate the context and spatial distribution of over-represented k-mers in yeast USRs.

Results: Single and double-stranded spatial distributions of most over-represented k-mers are highly non-random, and predominantly cluster into a small number of classes that are robust with respect to over-representation measures. Specifically, we show that the three most common distribution patterns can be related to DNA structure, function, and evolution and correspond to: (a) homologous ORF clusters associated with sharply localized distributions; (b) regulatory elements associated with a symmetric broad hill-shaped distribution in the 50–200 bp USR; and (c) runs of As, Ts, and ATs associated with a broad hill-shaped distribution also in the 50–200 bp USR, with extreme structural properties. Analysis of over-representation, homology, localization, and DNA structure are essential components of a general data-mining approach to finding biologically important k-mers in raw genomic DNA and understanding the 'lexicon' of regulatory regions.

Contact: hampson@ics.uci.edu; kibler@ics.uci.edu; pfbaldi@ics.uci.edu

1 INTRODUCTION

Over-represented k-mers in genomic DNA regions are often of particular biological interest. For example, over-represented k-mers in co-regulated families of genes are often associated with the DNA binding sites of transcription factors (van Helden *et al.*, 1998; Brazma *et al.*, 1998). In this case over-representation is measured by comparing the frequency of k-mers in the co-regulated genes to their frequency over all genes. However, over-representation can be compared between other sets and quantified in various ways. Here we introduce a new measure of over-representation, which compares a $k$mer's frequency to the frequency of its one-base-difference neighbors. Then we investigate the spatial distribution of over-represented k-mers in yeast Upstream Regions (USRs).

Somewhat surprisingly, in yeast USRs, almost all over-represented k-mers have distinctive distribution patterns. Distinctive distribution patterns result from a variety of biological factors, so k-mers with non-random localization patterns generally have other non-random properties as well. Understanding a string's distribution pattern is an important component in inferring its biological significance.

In this paper we focus on some of the most common spatial distribution patterns for over-represented k-mers in yeast. In particular, for $k = 8$ or 9 we show that the three most common distribution patterns correspond to: (a) homologous ORF clusters associated with sharply peaked distributions; (b) regulatory elements associated with a symmetric broad hill-shaped distribution in the 50–200 bp USR; and (c) runs of As, Ts, and ATs associated with a broad hill-shaped distribution also in the 50–200 bp USR, with extreme structural properties. This distribution is slightly asymmetric for runs of As and Ts, and trivially symmetric for runs of ATs. There is some overlap between the second and third patterns, but none with the first.

2 METHODS

Measuring over-representation is a statistical problem that depends on several considerations including: (a) selection of DNA regions; (b) inclusion of expression data; (c) motif representation; (d) counting methods; (e) background models; and (f) over-representation statistics.

*To whom correspondence should be addressed. Also at Department of Biological Chemistry, College of Medicine.

© Oxford University Press 2002

Sequence data

While over-representation analysis can be applied to any class of genomic DNA sequences, from whole genome to upstream or downstream regions of co-regulated ORF sets (Brazma *et al.*, 1998; van Helden *et al.*, 1998, 2000), here we focus primarily on the 500 bp USRs of each of the 6225 ORFs in the yeast genome taken from the Stanford data base (fttp://genome-ftp.standford.edu/yeast). For convenience, we often use the same label for the ORF and its USR. It is assumed that much of the control for each ORF's transcription rate resides in this region. Base positions in the USR are numbered positively with zero on the right, nearest the coding region.

Microarray data

The methods to be presented can be applied to raw genomic data. For comparison purposes, however, and because such data is becoming increasingly available, in some of the analysis we also use microarray data from which classes of co-regulated genes can be inferred using clustering methods (DeRisi *et al.*, 1997; Eisen *et al.*, 1998; Brown *et al.*, 2000; Hu *et al.*, 2000). We use the same data as in Hampson *et al.* (2000) derived by studying the oxidative stress response in yeast using Affimetrix Gene Chip microarray technology (Wodicka, 1997). In a typical experiment, the wildtype yeast strain YPH500 (Sikorski and Hieter, 1989) was used with three untreated controls grown at room temperature and two treated data sets, assayed independently. Oxidative stress treatment was given in the form of 0.4 mM of oxygen peroxide (H_2O_2) for 5, 10, and 20 min. GeneChip Expression Analysis v. 3.1 software was used to obtain the average difference values. All experiments were prepared using the polyA mRNA protocol.

Motif representation

In van Helden *et al.* (1998), binding sites are represented as exact k-mers while in Brazma *et al.* (1998) a restricted regular expression language is used to find them via an efficient algorithm for computing suffix trees. The most common recent representation for DNA binding sites is probability matrices (Bailey and Elkan, 1995; Chen *et al.*, 1995; Brown *et al.*, 2000; Hu *et al.*, 2000) or, more generally, hidden Markov Models (MMs; Baldi and Brunak, 2001). In order to make the search through the space of probability matrices tractable, most programs carry out some form of heuristic search, resulting in the well-known problems with hill-climbing algorithms (see also Pevzner and Sze, 2000; Pevzner, 2000 and references therein). We have developed a hill-climbing algorithm to optimize IUPAC motifs or weight matrices for over-representation (Hampson *et al.*, 2000; Kibler and Hampson, 2001), but the process of motif optimization is not pursued here since we only analyze exact k-mers.

Counts

The number of occurrences, $C0$, of each of the 4^k k-mers ($k \leqslant 9$) can be collected in a single pass through the data set. We count non-overlapping occurrences and report the results for $k = 7, 8$ or 9 only, although useful results are also obtained with neighboring values of k. It is well known that certain regulatory motifs are active regardless of the strand on which they occur and these are better detected if the counts on both strands are aggregated. Other motifs are strand-specific. Hence we compute both aggregated and single-strand counts.

Background model

Over-representation of a $k$mer must be assessed with respect to a statistical background model. We use MMs of order $l < k$ which can be measured on the data or some other reference set (Burge *et al.*, 1992; van Helden *et al.*, 2000; Bussemaker *et al.*, 2000; Baldi and Brunak, 2001) to produce an expected estimate $E_l(C0)$. As a point of general comparison, based on the average first-order nucleotide composition of yeast ($A \approx T \approx 30\%$, $C \approx G \approx 20\%$), the expected number of occurrences for a random 9mer in the data set is between 2 and 60. Some strings of interest are well in excess of this range, but many are not. We also introduce a new background model where we estimate the counts of each $k$mer as the average of $C1$, the summed count of all k-mers that differ in one position. This 'mean field' approximation can be extended to include multiple mismatches ($Cn, n > 1$), or insertions and deletions, although these directions are not further explored here.

Over-representation statistics

Several statistics can be used to detect significant over-representation with respect to background, such as ratio, log-likelihood, z-score, t-test, Poisson, and compound Poisson. Here we use ratios, z-scores, and binomial probabilities. Different combinations of background models and over-representation statistics can yield different rankings of k-mers with different task-specific tradeoffs. The localization results, however, are robust with respect to the different rankings because: (a) there is substantial overlap between the sets of top-ranked k-mers across the most efficient combinations; and (b) for all combinations, the three localization patterns we describe are dominant among the top-ranked k-mers, although the relative proportion of each pattern varies. Once robustness is established, the localization results are presented in terms of the ranking derived from the single-mismatch model with ratio statistics associated with reasonably large counts (above 10). There are $3k$ single-mismatch neighbors, hence we use the normalized ratio $3k \times C0/C1$ to assess over-representation. $C0$ can be computed in a single $O(N)$ pass, where N is the number of bases in

the set (here $N = 500 \times 6225 \approx 3 \times 10^6$). Once $C0$ has been computed for all k-mers, $C1$ can be computed in $3k$ steps for each of the 4^k k-mers. The total time complexity of computing $C0/C1$ for all k-mers is therefore $O(N) + O(k * 4^k)$. For $k \leqslant 9$, this takes only a few seconds on a workstation. Successive values of Cn, for increasing values of n, can also be computed in $3k \times 4^k$ steps as combinations of Cp's, $p < n$.

Spatial distribution

A distinctive spatial distribution pattern on one or both strands can provide supportive information regarding the structure, function, or evolution of a $k$mer. Regulatory motifs can often occur on either strand, but some distribution patterns are one-stranded or distinctly different on the two strands, so the distribution on each strand is shown separately. In practice, this is computed by searching the transcribed strand for both a string and its Reverse Complement (RC). The histograms associated with the location of a $k$mer are displayed back to back, facing up for the transcribed strand and down for the untranscribed strand.

Context analysis

A conserved context around an over-represented $k$mer on one or both strands can also provide useful information. Thus, once a particular $k$mer is selected, we study its context using both local and global alignments of the USRs in which it is found. Typical parameter values used in the alignment scoring function are: $match = 1$, $mismatch = -1$, $start\ delete = -2$, and $continue\ delete = -1$. Using those values, the average score of random USR pairs is -44, a score over 0 indicates some level of non-random homology, and a score over 50 virtually assures it (Hampson *et al.*, 2000). However, considerable local homology may exist without being apparent in the global homology score. Empirically, some USR pairs with local homology of over 200 bp have a negative global alignment score. For local alignments, we report the length of the longest highly homologous region, where 'highly' is defined as maintaining a positive alignment score over the length of the homologous region despite increasing the *mismatch* and *delete* values to -3. A larger value such as -5 would identify shorter, more highly homologous regions.

To further investigate a $k$mer's context, we also build a probability matrix for a window of $\pm m$ bases around it, as well as the corresponding consensus string. Conservation in the window can be assessed by computing at each position the relative entropy (Baldi and Brunak, 2001) $\mathrm{RE} = \sum_X P_X \log(P_X/Q_X)$ between the first-order background distribution Q measured over the entire data set and the observed distribution P over $X =$ A, T, G, C.

Table 1. Exact TRANSFAC matches for the 100 most frequent and 100 least frequent k-mers by three measures of over-representation versus 100 random k-mers. A string is counted as matching TRANSFAC if either the string or its RC exactly matches a yeast entry, or is a substring of a yeast entry

	$C0$		$C0/E_1(C0)$		$C0/C1$		
k	High	Low	High	Low	High	Low	Random
7	71	16	48	12	63	12	25
8	38	4	24	2	40	3	6
9	19	0	14	0	17	0	2

3 RESULTS

3.1 Robustness

We ranked all possible k-mers for $k = 7, 8$ and 9 using $C0$, the single-mismatch background model, as well as all MMs of orders 1 to $k - 1$, and different over-representations statistics including ratios, z-scores, and binomial probabilities (see Tables 1–3 for a sample of results). Table 1 gives the number of TRANSFAC matches using three representative measures. Yeast entries in TRANSFAC are AT-rich (about 56% AT), primarily due to runs of As and Ts, but also due to alternating As and Ts. Sorting on $C0$ alone yields strings with a bias toward high AT/GC ratios, including runs of As and Ts, and has a high TRANSFAC hit rate. These strings are biologically significant but, on the whole, are less interesting than over-represented strings with a more balanced AT/GC ratio. This problem can be addressed by using a low order (e.g. $l = 1$) background MM, but the ordering still has a noticeable bias toward strings containing long runs of As and Ts (Table 2). Eliminating the AT-rich bias actually reduces the number of TRANSFAC matches. $C0/C1$ reduces, but does not completely eliminate the bias for AT-rich strings or runs of As and Ts, but it identifies many additional interesting GC-rich strings with matches in TRANSFAC.

Markov models of higher orders (close to k) are also not optimal background models because their prediction is too close to the true value. This is obvious for models of order exactly k, but can also be seen for other high orders. For example, in a simulation, we generated 6000 500 bp-long strings using an approximate first order yeast distribution of $A = T = 30\%$ and $C = G = 20\%$. In each string, we inserted the 8mer ATGCCGTA. As a result, this 8mer is highly over-represented and is ranked at the top of the list by $C0$ or $C0/C1$ with $C0 = 6044$ and $C1 = 923$. Using a background MM of order $k = 6$ or 7, however, this motif cannot be distinguished from the random background with ratio or z-scores. Thus both low- and high-order MMs have drawbacks which are not present in the model we propose. Based on our experience, we recommend using MMs of order

Table 2. Top 20 9-mers sorted on $C0$ alone, $C0/E_1(C0)$, and $C0/C1$ (normalized value $= 27 \times C0/C1$)

$C0$	$C0$	$C0/E1$	$C0/C1$	$C0/E_1$	$C0$	$C0/E1$	$C0/C1$	$C0/C1$	$C0$	$C0/E1$	$C0/C1$
TTTTTTTTT	1320	15.01	2.87	TTTTTTTTC	994	18.77	4.08	GGCTAAGCG	25	6.50	7.67
AAAAAAAAA	1250	12.64	2.80	GCGATGAGC	67	17.00	6.62	GCGATGAGC	67	17.00	6.62
TTTTTTTTC	994	18.77	4.08	CGCGCGCGC	15	16.00	5.96	TCGGCGGCT	35	12.00	6.52
GAAAAAAAA	906	15.91	3.94	GAAAAAAAA	906	15.91	3.94	GACTCCCCG	13	4.67	6.50
CTTTTTTTT	805	15.21	3.40	CTTTTTTTT	805	15.21	3.41	ACGCGCGCG	15	8.00	6.32
ATTTTTTTT	784	8.82	3.11	TTTTTTTTT	1320	15.01	2.88	CGCGCGCGC	15	16.00	5.96
AAAAAAAAT	734	7.58	3.04	GCGCGCGCC	14	15.00	4.79	CCTCGAGGA	46	11.75	5.89
AAAAAAAAG	715	12.56	3.21	TTTCTTTTT	689	13.02	3.27	TCCTCGAGG	44	11.25	5.63
TTTCTTTTT	689	13.02	3.27	GCTCATCGC	51	13.00	5.26	CGATGAGCT	70	10.14	5.61
TTTTTTTCT	668	12.62	3.01	AAAAAAAAA	1250	12.64	2.80	CGAGGGTCC	16	5.67	5.61
AAAAAGAAA	655	11.51	3.17	TTTTTTTCT	668	12.62	3.01	CGGGGTTCG	13	4.67	5.57
TTTTCTTTT	629	11.89	2.88	AAAAAAAAG	715	12.56	3.21	TAGCCGCCC	26	9.00	5.53
AAAAGAAAA	620	10.89	2.86	TCGGCGGCT	35	12.00	6.52	GGATTCCTA	42	3.58	5.48
TCTTTTTTT	608	11.49	2.84	TTTTCTTTT	629	11.89	2.88	GGAGACCGG	14	5.00	5.48
TTTTTTCTT	585	11.06	2.74	CCTCGAGGA	46	11.75	5.89	ACCACACCC	36	7.40	5.46
ATATATATA	582	6.20	3.83	AAAAAGAAA	655	11.51	3.17	TTAGCCGCC	33	8.50	5.30
AGAAAAAAA	576	10.12	2.69	TCTTTTTTT	608	11.49	2.84	ACGAGGGTC	17	4.50	5.28
TTTTTCTTT	570	10.77	2.64	TCCTCGAGG	44	11.25	5.63	CATCTCATC	90	7.58	5.27
TATATATAT	569	6.13	3.94	TTTTTTCTT	585	11.06	2.75	GCTCATCGC	51	13.00	5.26
AAGAAAAAA	561	9.86	2.69	AAAAGAAAA	620	10.89	2.86	CCCCACGGA	15	5.33	5.26

Table 3. Comparison of top 20 8-mers sorted using different background models and statistics: background MMs of order 4 and 5, with z-scores and binomial probabilities, versus $C0/C1$ ('*' indicates that the 8mer is found among the top 20 identified by $C0/C1$)

MM4 z-score	MM5 z-score	MM4 binomial	MM5 binomial	$C0/C1$
GCGATGAG*	GCGATGAG*	AAAAGAAA	GCGATGAG*	GCGATGAG
TATATATA*	GTTACCCG*	ATATATAT*	AAAAGAAA	TAGCCGCC
ATATATAT*	CGGGTAAC*	ATACATAT	GTTACCCG*	CCGGGTAA
GTTACCCG*	CTCATCGC*	TATATATA*	CTCATCGC*	GTTACCCG
CTCATCGC*	CCGGGTAA*	TTTTTTTC*	TTTCTTTT*	CTCATCGC
CCGGGTAA*	TAGCCGCC*	TTTCTTTT	CGGGTAAC*	GGCGGCTA
CGGGTAAC*	AAAAGAAA	GAAAAAAA*	CCGGGTAA*	TTTTTTTC
TTACCCGG*	ATTACCCG*	GATGAGCT*	ATTACCCG*	CGATGAGC
GATGAGCT*	TTACCCGG*	GTTACCCG*	GATGAGCT*	CGGGTAAC
CGATGAGC*	GATGAGCT*	GCGATGAG*	TTACCCGG*	TTACCCGG
ATTACCCG*	CGATGAGC*	CATATATA	CGGGTAAT*	ATTACCCG
ATATGTAT	CGGGTAAT*	CTCATCGC*	CGATGAGC*	TCCGGGTA
ATACATAT	TTTCTTTT	CCGGGTAA*	TAGCCGCC*	GAAAAAAA
TTTTTTTC*	ATGAGATG	ATATGTAT	AAACAAAA	GATGAGCT
CATATATA	TCACGTGA	TATACATA	ATGAGATG	CCTCGAGG
TATACATA	TCCGGGTA*	CGGGTAAC*	GATGAGAT	TCGGCGGC
TAGCCGCC*	GATGAGAT	CGATGAGC*	TTTTCTTT	TATATATA
CGGGTAAT*	AAACAAAA	TTACCCGG*	TGAAAAAT	CGGGTAAT
AAAAGAAA	ATCTCATC	ATTACCCG*	TAATATTA	ATATATAT
TATGTATA	GTCACGTG	TATGTATA	ATACATAT	CACGTGAC

close to $k/2$ as a rule of thumb as well as the single-mismatch model, or one of its variants. Indeed, different models and statistics have different biases and may be used selectively for different tasks, or simultaneously for comparison purposes. The one-mismatch model, for instance, is particularly suited for detecting k-mers where the exact pattern is over-represented compared to the density of patterns in its immediate vicinity. Several variants, such as using $(C0 + C1)/C2$, are possible. It is also worth noting that, unlike the case of MMs of any order, for any given $k$mer none of its occurrences counted in $C0$ are used to compute the expected background value $C1/3k$.

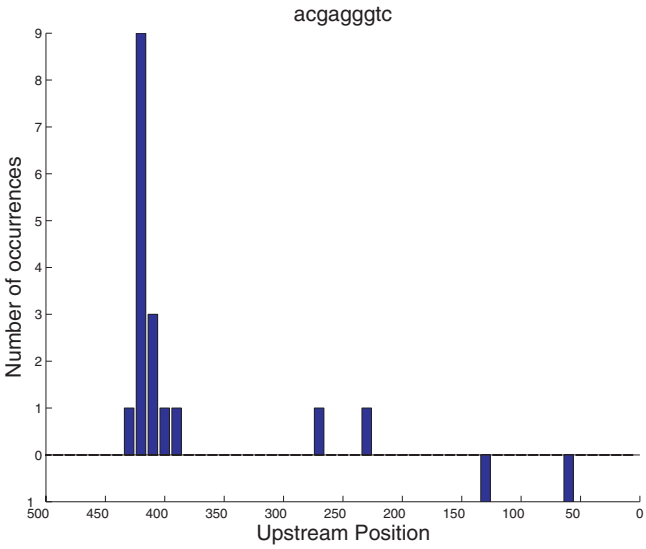
In terms of over-representation statistics, small counts

Table 4. Pattern frequencies among the top 20 8- and 9-mers, sorted on $C0/C1$, with or without the reverse complement. Homologous USRs due to transposons listed in Kim *et al.* (1998) are counted separately, but are not analyzed further in this paper. Numbers in parentheses count the patterns with exact matches in TRANSFAC

k	8	8 + RC	9	9 + RC
Homologous ORFs	1	0	7	3
Transposons	1	2	6	2
Motifs	11 (4)	8 (5)	4 (0)	9 (3)
Poly A/T	4	6	0	1
Other	3	4	3	5

can be problematic for ratios (e.g. 1 versus 2 is different from 10 versus 20), z-scores, or Gaussian approximation to the corresponding binomial model. Binomial probabilities require more computations but have less of these problems. On the other hand, z-scores, Gaussian approximations, and binomial probabilities always rank an observed/expected pattern of $20n/10n$ higher than a pattern of $4n/n$, no matter how large the value of n, a feature that is not always desirable. In our experience, good results can be obtained with ratios as long as the counts are not too small. MMs of order $k/2$ together with binomial probabilities may yield slightly better results if the goal is to find transcription factor binding sites, especially those with high counts. But there are also plenty of interesting strings which most likely are not transcription factor binding sites, which are over-represented but with counts in a lower range than the most common regulatory motifs. These are likely to be easier to detect with a $C0/C1$ ratio. In any case, while the choice of a particular background model and statistics impacts the ranking, it does not impact the localization results. For example, 14 of the top 20 8-mers obtained by z-score with a MM of order 4 are in the top 20 8-mers obtained by $C0/C1$. Likewise, 13 of the top 20 8-mers obtained using binomial probabilities with a MM of order 4 are in the top 20 8-mers obtained by $C0/C1$ (Table 3). Sixteen of the top 20 8-mers obtained with the MM of order 4 with either statistical measure are found in the top 50 obtained by $C0/C1$. Not only is there a high degree of overlap between the top ranked k-mers of each list, but the great majority of the k-mers at the top of each list satisfies one of the three localization patterns to be described below (Table 4). In what follows, the main localization results are reported using $k = 9$, and the $3k \times C0/C1$ statistics with $C0 > 10$. In fact, the latter constraint is automatically satisfied by all top-ranked k-mers.

3.2 Sharply localized distributions: homologous ORFs



9mer	$C0$	$C1$	$27 \times C0/C1$
ACGAGGGTC	17	87	5.28
GACCCTCGT	2	95	0.57

Fig. 1. Distribution and counts for ACGAGGGTC and its RC.

3.2.1 One-stranded localization. If 9-mers are sorted on $C0/C1$ without including the RC, a distinct and easily explained pattern of localization is observed for many k-mers. For example, in Figure 1, the distribution of the string ACGAGGGTC and its RC GACCCTCGT is shown, along with their $C0$ and $C1$ counts and normalized $C0/C1$ ratio.

A tightly localized distribution on one strand only might result from a group of highly homologous ORFs associated with a series of duplication events. A single large duplication (Wolfe and Shields, 1997) in itself can only explain pairs of homologous ORFs, so large groups of homologous USRs can only be explained by additional duplication events, such as those resulting from shared transposable elements (Kim *et al.*, 1998). At least six of the top 20 $C0/C1$ k-mers in Table 2 are associated with transposon long terminal repeats but the homologous ORF families considered in this paper are not associated with any of the known transposons contained in the list at http://www.public.iastate.edu/∼voytas (Kim *et al.*, 1998). Seven of the top 20 $C0/C1$ 9-mers in Table 2 belong to large homologous ORF families that are not associated with transposons (Table 4).

As expected from homologous ORFs, the local context surrounding the string is indeed partially conserved (Table 5). If the set of ORFs containing a highly localized

Table 5. Base frequencies and Relative Entropy (RE) over the 17 positive-strand occurrences of ACGAGGGTC, context of five on each end. Resulting consensus sequence: TCTCG ACGAGGGTC CAAAT

A	0	0	0	0	29	100	0	0	100	0	0	0	0	0	5	64	64	70	17
T	88	11	82	5	64	0	0	0	0	0	0	0	100	0	5	17	17	5	64
G	0	5	5	82	0	0	0	100	0	100	100	100	0	0	0	5	11	17	5
C	11	82	11	11	5	0	100	0	0	0	0	0	0	100	88	11	5	5	11
RE	0.9	1.0	0.7	1.1	0.4	1.1	1.7	1.7	1.1	1.7	1.7	1.7	1.2	1.7	1.2	0.2	0.2	0.4	0.2

over-represented string are aligned, they generally fall into one or two homologous sets, plus a few ORFs that are not strongly homologous to anything in the set. For example, local and global alignment scores for the USRs of the 19 ORFs containing the above string (or its RC) against the first ORF in the set are given in Table 6. Of the 19 ORFs, 13 fall in one homologous family, although the amount of global and local homology varies over a sizable range. All of the 13 homologous ORFs are annotated in the Stanford database as either: 'strong similarity to subtelomeric encoded proteins' or 'strong similarity to members of the Srp1/Tip1p family' indicating that their coding regions are also homologous and of similar function. Pairs with a global alignment score in the 300s but a local score near 500 are moderately homologous over the entire USR rather than highly homologous over a portion of it. The string's location in the 13 homologous ORFs corresponds to the spike in the distribution of the whole set. Such highly concentrated unimodal distributions are characterized by their small location variance and therefore location variance can be used to detect and further study them. Strings with low variance usually have elevated $C0/C1$ ratios, although the converse is not generally true.

A string may fall into more than one non-overlapping cluster of homologous ORFs (Figure 2). Global alignment shows two clusters based on alignment with the first and second ORFs (Table 7). These two sets correspond to the two closely spaced spikes around position 370. There is very little homology between the clusters beyond the shared 9mer. The fact that a specific 9mer is localized in approximately the same position in two very different ORF families may be due to random chance, but might indicate functional significance (see also Appendix A).

3.2.2 Two-stranded localization: divergent ORFs. Another variant of highly localized distribution occurs when both the string and its RC are highly localized (Figure 3). In such cases, the ORFs containing the RC generally form a separate homology group from those containing the string itself. In this example, alignment with YCR104W (which contains the string on the direct strand) produces a homology set of nine ORFs, and alignment with YBL108W (which contains the RC) produces a separate homology set of six ORFs (Table 8).

Table 6. Alignment scores of 19 USRs containing the sequence: ACGAGGGTC, or its reverse complement, to the first one of them (YALO68C). G = global, L = local, R = reverse complement, C = cluster number

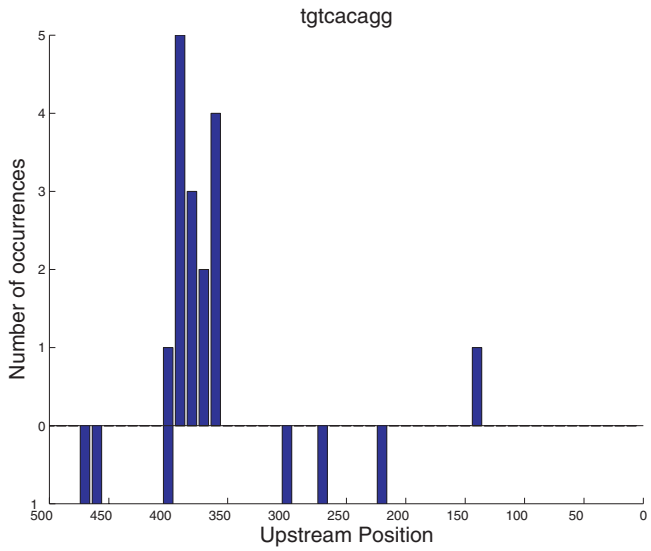
			G1	L1	C
1	YAL068C		500	500	1
2	YCR104W		374	500	1
3	YDR542W		145	34	1
4	YER109C		−58	11	
5	YGL261C		383	500	1
6	YGR130C	R	−43	9	
7	YHL046C		301	438	1
8	YIL176C		380	500	1
9	YIR041W		397	498	1
10	YJL223C		380	500	1
11	YKL199C	R	−60	8	
12	YKL224C		397	498	1
13	YKR086W		−57	10	
14	YLL025W		33	23	1
15	YLL064C		351	474	1
16	YLR048W		−52	10	
17	YLR461W		381	499	1
18	YNR076W		341	499	1
19	YOL028C		−63	10	

The 20-base consensus contexts of the string (S) and its RC (R) show almost perfect local homology:

```
S=AAAGATGAGATATGGAGGAT[-]GCTAAATGAGCATCTGTTAA
R=AAAGATGAGATATGGAGAAT[-]TCTAAATGAGCATCTGTTAA
[-]=ATGTGAGGT
```

which extends to a broader local region. The local alignment scores of all the ORFs (with the RC of the first and third ORFs in Table 8) show an homologous region roughly 200 bases long. ORF number 10 is an exception with a homologous region of roughly 300 bases. Extended local RC homology suggests that one set might be globally homologous to the RC of the other, however this is not the case.

This particular distribution pattern is associated with divergently transcribed ORFs that have frequently replicated together, since it contains a substantial fraction of pairs of ORFs that are physically adjacent but on different strands of DNA (e.g. YCR103C and YCR104W) with overlapping



9mer	$C0$	$C1$	$27 \times C0/C1$
TGTCACAGG	16	194	2.23
CCTGTGACA	7	200	0.95

Fig. 2. Distribution and counts for TGTCACAGG and its RC.

Table 7. Alignment scores of 23 USRs containing the sequence TGTCACAGG, or its reverse complement, to the first and second ones (YALO68C and YBR302C). G = global, L = local, R = reverse complement, C = cluster number

			G1	G2	L1	L2	C
1	YAL068C		500	−49	500	9	1
2	YBR302C		−49	500	9	500	2
3	YCR027C	R	−40	−56	9	8	
4	YDL248W		−46	466	14	495	2
5	YDR519W	R	−33	−30	9	10	
6	YEL031W		−38	−34	10	11	
7	YER042W	R	−35	−42	9	9	
8	YFL042C	R	−45	−50	10	9	
9	YFL062W		−46	393	11	493	2
10	YGL055W	R	−42	−48	9	8	
11	YGL261C		383	−39	500	10	1
12	YGR295C		−43	350	11	497	2
13	YHL034C		−47	−34	9	9	
14	YHL048W		−51	482	9	500	2
15	YIR041W		397	−40	498	10	1
16	YJR161C		−49	490	9	500	2
17	YKL224C		397	−46	498	10	1
18	YLL064C		351	−22	474	9	1
19	YML132W		−49	500	9	500	2
20	YMR014W	R	−34	−40	8	13	
21	YNL336W		−47	488	9	499	2
22	YPL222W		−46	−45	11	13	
23	YPR136C	R	−54	−41	12	8	

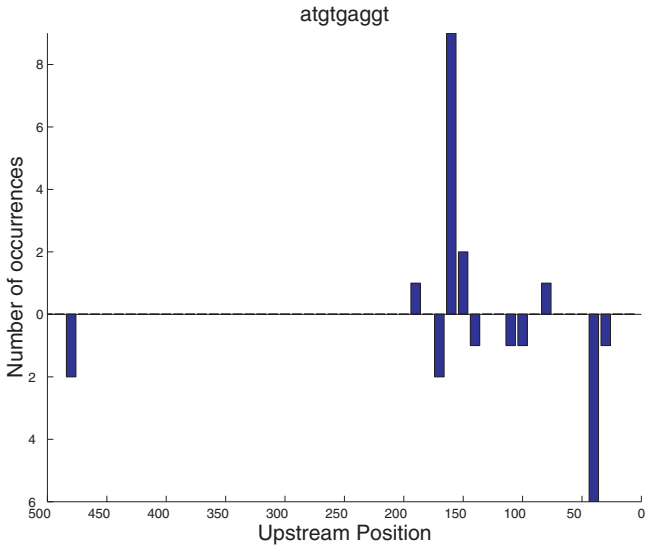
USRs. A distance of 500 between the divergent ORFs (that is a distance of 500 between their '0' points) would produce perfect RC homology in the current 500-base data set. The divergent pairs, associated with the first cluster in Table 8, have a distance of approximately 200, resulting in a local RC homology of approximately 200 bases. One possibility for the homologous but unpaired ORFs is that they replicated as intact pairs but sufficient mutations accumulated so that one member of the pair is decayed and no longer recognizable as an ORF. For example, YBL108W does not have a divergent partner, but the divergent 'empty' region immediately upstream of it is in fact homologous to YCR104W. Thus, an homologous pseudogene can be assigned to that region.

To summarize the homology results, sorting 9-mers on $C0/C1$ without the RC identifies a number of strings with tightly localized distribution patterns. In general, these strings are the most conserved portion of larger homology regions between ORFs and can be further investigated by specifically looking for strings with similar localized distribution patterns. This is achieved by sorting on location variance. Localized strings are used to identify groups of homologous ORFs, which are then further analyzed for conserved regions. Once an initial homology group is identified, additional group members are extracted by aligning a group member against the entire data set. Distributions with more than one spike often result in separate homology groups for each spike. Distribution patterns in which both the string and its RC are localized generally define separate homology groups resulting from divergently transcribed ORFs.

3.3 Broad symmetric distribution: regulatory motifs

A second frequent class of spatial distribution patterns for over-represented k-mers is shown in Figure 4 and consists of a broad, fairly symmetric, hill localized around 50 to 200 bps. This corresponds to a preferred region observed in previous work (Brazma *et al.*, 1998; Hampson *et al.*, 2000; Hughes *et al.*, 2000). Including the RC in the calculation of $C0/C1$ helps identify symmetric distributions but it is not a necessity for finding them. For 8-mers, this is the most common distribution pattern for strings with high $C0/C1$. For 9-mers it is a common pattern if both the string and its RC are included in the $C0/C1$ calculation. Close inspection of the individual strings with this distribution pattern indicates that many of them result from a small set of longer, degenerate motifs.

A similar distribution arises for another string (Figure 5) which will be considered in more detail. Most of the general conclusions drawn about it are applicable to the pre-



9mer	$C0$	$C1$	$27 \times C0/C1$
ATGTGAGGT	13	206	1.70
ACCTCACAT	14	244	1.55

Fig. 3. Distribution and counts for ATGTGAGGT and its RC.

vious string. In fact, the two have a strong tendency to co-occur in the same USRs. Both have been previously identified based on other extraction mechanisms (Hampson *et al.*, 2000; Hughes *et al.*, 2000). There is no evidence of global or extended local homology between the ORFs containing the string, but there is a limited amount of local homology around the shared string as measured by the consensus context:

```
S=AATTTTTTTTTATTAATTTT[-]TTAAAAAAAAAATTAAAAAA
R=TATTTAATTTTTTAAATTTT[-]TTAAAAAAAAAAATAATAAA
[-]=GCGATGAGC
```

The consensus context indicates some local conservation between the direct and RC strands. However, it over-states the case since it does not indicate the actual amount of variability. For example, while the 5-base consensus context is identical for the string and its RC in the above example, it is actually quite variable on a case-by-case basis (Table 9). In spite of the variability at each position, the two probability tables are actually similar. This is because an abnormally large number of occurrences are in the context of divergent ORF pairs. There are 1343 divergent pairs that overlap in their 500 bp region, with an average overlap of approximately 290, so the probability that any given occurrence of a $k$mer will also be counted as its RC in a divergent partner is approximately $((2 \times 1343)/6225) \times (290/500) = 0.25$, or 0.27 if measured empirically for random 2, 3 and 4-mers. By this argument, approximately $(67 + 51)/2 \times 0.27 = 16$ in each set should be due to divergent ORF pairs, when in fact 36 are. This produces a high degree of symmetry in the two probability tables. The biological significance of this association with divergent pairs is unknown, but it is probably not a coincidence that many of the divergent pairs containing this string overlap in a way that preserves the preferred location (approximately 120) for both the string and its RC.

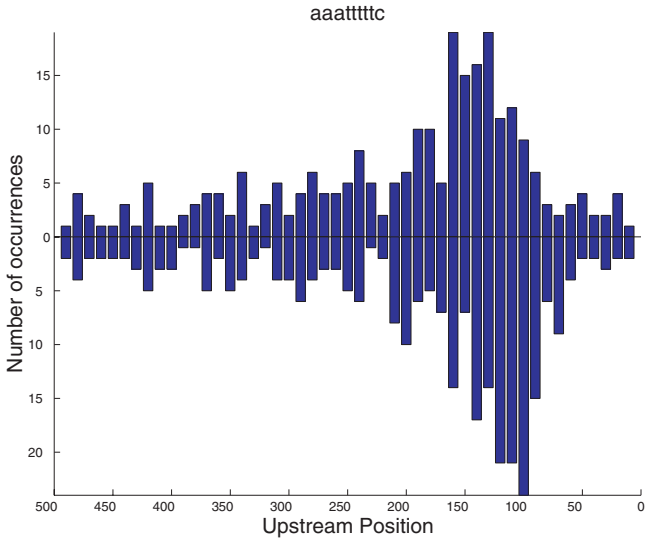
Strings with broad distribution patterns are not adequately identified by sorting on location variance, but it is possible to devise a different localization filter by looking at the ratio of the counts in the 50–200 bp region to the counts in the 500 bp USRs, in isolation or in combination with $C0/C1$, or by using a similar chi-square test. Most of the top-ranked strings found by this tuned ratio, applied to strings plus their RC with at least 40 counts in the 50–200 bp region, have: (a) a similar symmetric distribution for the string and its RC; (b) an elevated $C0/C1$ ratio; and (c) an unexpectedly high number of divergent pairs. Many strings identified this way show a strong correlation

Table 8. Alignment scores of 27 USRs containing the sequence ATGTGAGGT, or its reverse complement, to the first and third ones (YBL108W and YCR104W). G = global, L = local, R = reverse complement, C = cluster number. GR1 and GR3 do not show any global homology (not shown)

			G1	G3	LR1	LR3	C
1	YBL108W	R	500	−72	14	208	1
2	YCR103C	R	392	−63	10	208	1
3	YCR104W		−72	500	208	12	2
4	YCRX09C		−82	−66	11	9	
5	YDR487C		−84	−91	12	8	
6	YGL260W	R	433	−75	14	204	1
7	YGL261C		−70	397	204	11	2
8	YHL022C	R	−57	−78	12	11	
9	YIL150C	R	−79	−62	8	14	
10	YIL174W	R	139	−71	14	310	1
11	YIL176C		−71	413	204	11	2
12	YIR040C	R	353	−66	14	205	1
13	YIR041W		−67	398	205	11	2
14	YJL223C		−71	413	204	11	2
15	YJR082C	R	−74	−64	11	12	
16	YKL223W	R	347	−67	14	202	1
17	YKL224C		−63	390	202	11	2
18	YLL064C		−68	446	208	10	2
19	YLR461W		−77	401	204	12	2
20	YMR214W	R	−80	−72	9	9	
21	YNR076W		−59	437	205	10	2
22	YOL121C		−87	−61	10	19	
23	YOR190W	R	−73	−97	10	9	
24	YOR295W	R	−75	−85	10	9	
25	YPL267W	R	−71	−69	8	10	
26	YPR162C		−79	−70	10	9	
27	YPR185W	R	−60	−75	11	11	

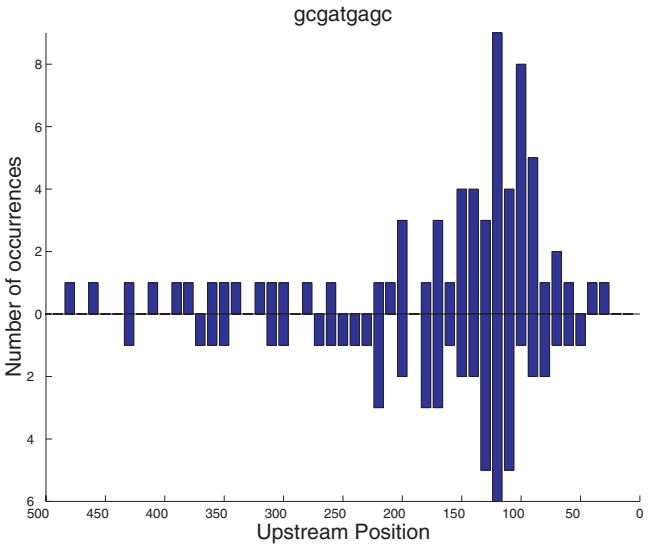
Table 9. Top half: base frequencies and relative entropy for the 67 positive-strand occurrences of GCGATGAGC, context of five on each end. Resulting consensus sequence: ATTTT GCGATGAGC TTAAA. Bottom half: same for 51 RC occurrences. Resulting consensus sequence: ATTTT GCGATGAGC TTAAA

A	44	32	31	40	16	0	0	0	100	0	0	100	0	0	20	22	47	43	46
T	28	38	43	41	53	0	0	0	0	100	0	0	0	0	70	41	19	23	17
G	13	10	14	8	17	100	0	100	0	0	100	0	100	0	2	28	10	28	22
C	13	17	10	8	11	0	100	0	0	0	0	0	0	100	5	7	22	4	13
RE	0.0	0.0	0.0	0.1	0.1	1.7	1.7	1.7	1.1	1.2	1.7	1.1	1.7	1.7	0.4	0.1	0.1	0.1	0.1
A	41	31	31	31	13	0	0	0	100	0	0	100	0	0	19	23	43	43	37
T	37	41	43	45	62	0	0	0	0	100	0	0	0	0	70	45	19	23	21
G	1	11	11	11	9	100	0	100	0	0	100	0	100	0	0	25	11	27	23
C	19	15	13	11	13	0	100	0	0	0	0	0	0	100	9	5	25	5	17
RE	0.1	0.0	0.0	0.1	0.2	1.7	1.7	1.7	1.1	1.2	1.7	1.1	1.7	1.7	0.4	0.1	0.1	0.1	0.0



9mer	$C0$	$C1$	$27 \times C0/C1$
AAATTTTTC	254	2212	3.10
GAAAAATTT	288	2334	3.33

Fig. 4. Distribution and counts for AAATTTTTC and its RC.



9mer	$C0$	$C1$	$27 \times C0/C1$
GCGATGAGC	67	273	6.62
GCTCATCGC	51	262	5.26

Fig. 5. Distribution and counts for GCGATGAGC and its RC.

with expression regulation during oxidative stress so, as in the case of homologous ORF families, the set of ORFs containing the string can be analyzed separately as a co-regulated ORF family (Appendix B). A number of strings are over-represented in this set besides the string used to define it. These additional conserved strings are candidate regulatory motifs.

The results in Appendix B are with respect to oxidative stress, but strings identified by $C0/C1$ presumably also correlate with other gene expression patterns resulting from other experimental treatments. A good example is TTACCCGGA ($C0 = 45, C1 = 298$), and its RC TCCGGGTAA ($C0 = 50, C1 = 298$), with normalized $C0/C1$ ratios of about 5. Like the previous string, this is an identified motif (Brazma *et al.*, 1998; Hughes *et al.*, 2000) that has a symmetric distribution, little base conservation in its consensus context, and occurs in divergent ORFs more frequently than expected. But unlike the previous strings, it is not correlated with a change in expression during oxidative stress.

To summarize the regulatory motif results, sorting 8 and 9-mers on $C0/C1$ identifies a number of strings with

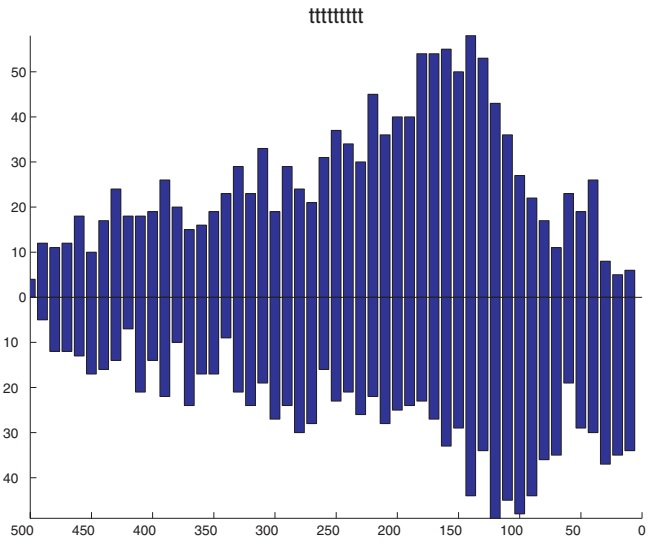
a broad, approximately symmetric, localization pattern in the 50–200 bp USR. These strings do not participate in a broader homology region. Many of them appear to result from a small number of degenerate motifs. The strings occur in the context of divergent ORFs more frequently than expected and tend to co-occur in the same USRs. Many of these strings are correlated with down regulation during oxidative stress, or simply with elevated expression during normal growth conditions. The effect may increase with the number of occurrences per ORF. Sorting on on/off, down/up, or the fraction of counts in the 50–200 region, preferentially identifies such strings.

3.4 Broad asymmetric distribution: strings of Ts and As

Ts and As are over-represented in the USR with frequencies of approximately A $\approx$ T $\approx$ 30% and C $\approx$ G $\approx$ 20%. However, even within this context, three types of AT-rich *k*-mers stand out because of their extreme over-representation and distinctive localization: long strings of Ts possibly interrupted by a single C, long strings of As possibly interrupted by a single G, and alternating Ts and As. The first two are RCs of each other and the third is its own RC. Similar sequences were reported in Hughes *et al.* (2000), and runs of Ts, As, and alternating TAs are also strongly over-represented with a broad hill-shaped spatial distribution in the region down-stream of ORFs (van Helden *et al.*, 2000). Interestingly, the dinucleotide TA is under-represented across a wide range of organisms, including yeast (Burge *et al.*, 1992; Karlin and Mrazek, 1997).

These strings are of special interest since they are the most frequent strings in the USR. Consequently, strings of this type can be identified simply by sorting on $C0$. The two strings consisting of all Ts and all As are always the first and second on the list when strings are sorted on $C0$ alone ($3 < k \leqslant 10$), and runs of Ts containing a single C and runs of As containing a single G occupy many of the other top spots (Table 2).

Runs of Ts (possibly containing a C) and runs of As (possibly containing a G) have a distinct, somewhat asymmetric distribution pattern (Figure 6). Runs of Ts and As are highly over-represented: for 9-mers, 1320 and 1250 versus an expected 60 based on the first-order frequency of Ts and As. They do not occur in the context of large-scale homology or in an unexpected number of divergent ORF pairs. They are not strongly correlated with changes in expression. However, they do show some correlation with high on/off values (Appendix B), and, conversely, sorting for high on/off identifies strings with runs of Ts and As. One of the strings considered in the previous section, AAATTTTTC, can be viewed from this perspective: it contains runs of both Ts and As, has a slightly asymmetric distribution much like longer runs of Ts/As, and has a high



9mer	$C0$	$C1$	$27 \times C0/C1$
TTTTTTTTT	1320	12 397	2.87
AAAAAAAAA	1250	12 060	2.80

Fig. 6. Distribution and counts for TTTTTTTTT and its RC.

on/off value.

The tendency for runs of Ts and As is reflected in the probability of seeing a T or A based on what precedes it (Table 11). The more Ts that precede a location, the more likely it is to be another T. In addition, a run of Ts is more likely to be followed by a C than a G. Similar results are obtained for runs of As. Cs and Gs show a quite different pattern. This shows that Ts and As tend to clump together, but does not explain the observed localization patterns. One possible explanation is that the distribution of runs of Ts and As simply results from the underlying distribution of the individual bases. That is, if individual Ts are denser in a particular region of the USR, runs of Ts would also be expected to be denser in that region. Based on the nucleotide probabilities at each of the 500 positions in the USR, the expected distribution pattern of any string can be computed. For $k \leqslant 4$, the expected and observed are close for runs of Ts, but for longer runs the observed is increasingly in excess of the expected. Thus, rather than the frequency of individual Ts determining the distribution of runs of Ts, the converse is more likely. A similar situation occurs with runs of As. The preference for interrupting a run of Ts with a C can be seen in the local context of the string TTTTCTTTT (Table 10). There is only partial base conservation, but the enhanced probability of Ts and Cs is still apparent.

Alternating Ts and As show a similar distribution

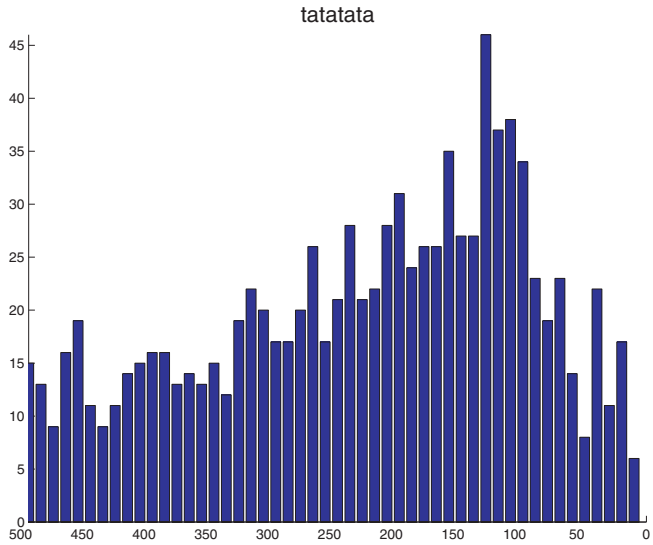
Table 10. Base frequencies and relative entropy for 629 positive-strand occurrences of TTTTCTTTT, context of five on each end. Resulting consensus sequence: TTTTT TTTTCTTTT TTTTT

A	21	23	19	22	14	0	0	0	0	0	0	0	0	0	8	14	18	16	20
T	45	42	48	46	53	100	100	100	100	0	100	100	100	100	58	48	50	49	48
G	12	13	11	10	9	0	0	0	0	0	0	0	0	0	11	10	10	11	12
C	20	20	20	21	21	0	0	0	0	100	0	0	0	0	20	25	20	22	18
RE	0.1	0.0	0.1	0.1	0.2	1.2	1.2	1.2	1.2	1.7	1.2	1.2	1.2	1.2	0.2	0.1	0.1	0.1	0.1

Table 11. Probability of extending a preceding run of As, Cs, Gs, and Ts

	Prob A	Prob T	Prob G	Prob C
Min # of prec. Ts				
0	0.3166	0.3128	0.1834	0.1872
1	0.2602	0.3662	0.1858	0.1882
2	0.2107	0.4047	0.1823	0.2031
3	0.1933	0.4394	0.1604	0.2083
4	0.1695	0.4753	0.1443	0.2124
5	0.1442	0.5360	0.1199	0.2016
6	0.1168	0.6082	0.0995	0.1769
7	0.0965	0.6453	0.0852	0.1748
Min # of prec. As				
0	0.3166	0.3128	0.1834	0.1872
1	0.3643	0.2806	0.1858	0.1684
2	0.4000	0.2448	0.1970	0.1568
3	0.4320	0.2284	0.1927	0.1435
4	0.4646	0.2135	0.1914	0.1265
5	0.5239	0.1983	0.1721	0.1029
6	0.6040	0.1659	0.1501	0.0776
7	0.6448	0.1492	0.1385	0.0655
Min # of prec. Gs				
0	0.3166	0.3128	0.1834	0.1872
1	0.3167	0.2784	0.1946	0.2106
2	0.3186	0.2770	0.1869	0.2181
3	0.3141	0.2809	0.1860	0.2195
4	0.3274	0.2504	0.1929	0.2300
5	0.3281	0.2142	0.2373	0.2197
6	0.3493	0.1603	0.3378	0.1489
7	0.3057	0.1189	0.4755	0.1019
Min # of prec. Cs				
0	0.3166	0.3128	0.1834	0.1872
1	0.3304	0.3116	0.1639	0.1942
2	0.3312	0.3144	0.1613	0.1930
3	0.3108	0.3385	0.1598	0.1922
4	0.3069	0.3294	0.1559	0.2083
5	0.2866	0.3097	0.1479	0.2566
6	0.2302	0.3611	0.1128	0.2979
7	0.1822	0.3036	0.0607	0.4555

(Figure 7). It is perfectly symmetric however, since it is RC invariant. The peak appears to fall between the peaks of runs of Ts and As. The consensus context shows that the string occurs in the context of longer runs of TA alternation.



8mer	$C0$	$C1$	$27 \times C0/C1$
TATATATA	1003	7710	3.51

Fig. 7. Distribution and counts for TATATATA (equal to its RC).

```
S=TATATATATATATATATATA[-]TATATATATATATATATATA
[-]=TATATATA
```

The 5-base context is TA-rich and shows the preference for TA alternation (Table 12). Global alignment of ORFs containing the string shows no evidence of global homology, and only limited local homology based on the extended region of TA alternation.

Runs of As, Ts, and ATs have very unusual structural properties. Homopolymeric dA : dT tracts are known from x-ray crystallography to be straight and rigid (Nelson *et al.*, 1987). This can also be ascertained using a number of dinucleotide or trinucleotide structural scales (Baldi *et al.*, 1999; Baldi and Baisnée, 2000), ranging from DNAse I bendability (Brukner *et al.*, 1995), to propeller twist angle (Hassan and Calladine, 1996), to protein deformability (Olson *et al.*, 1998). Such regions

of DNA are unlikely to bend easily and probably are bad candidates for nucleosome positioning when k is large. Studies in two different yeast species have shown that the homopolymeric elements destabilize nucleosomes and thereby facilitate the access of transcription factors bound nearby (Iyer and Struhl, 1995; Zhu and Thiele, 1996). A single G (resp. C) in a run of A's (resp. T's) preserves the purine (resp. pyrimidine) tract and is unlikely to modify the stiffness properties. In contrast, the triplet ATA/TAT, characteristic of the TATA box, is highly flexible according to the bendability scale and consistently with experimental results (Parvin *et al.*, 1995; Starr *et al.*, 1995; Grove *et al.*, 1996). Runs of alternating ATs can be shown to have maximal cumulative bendability (Baldi and Baisnée, 2000) and are likely to be associated with particularly flexible stretches of DNA.

4 DISCUSSION

Regulatory motifs can be identified by looking at over-represented k-mers in the regions surrounding classes of co-regulated genes derived, for instance, from microarray expression data or literature searches (van Helden *et al.*, 1998). Gene expression data, however, is not always available and can be highly variable (Lee *et al.*, 2000; Hegde *et al.*, 2000; Long *et al.*, 2001). Moreover, literature search is time-consuming if manual, and unreliable if automated. Furthermore, significant over-represented k-mers are not always associated with regulatory motifs.

We have developed an effective over-representation method to identify biologically important k-mers in both raw genomic sequences and filtered data. The method is efficient with run-time linear in the data. The method, and its possible generalization (e.g. $(C0 + C1)/C2$)) relies on a single-mismatch background model that can overcome the limitations of other methods that work well for highly conserved motifs but can degrade as variability increases. There is no reason to expect that most interesting biological sequences are over-represented against single mismatches, but the converse appears to be true: strings that are over-represented against single mismatches do appear to be biologically interesting and these are worth investigating for their own sake. We have applied over-representation methods to yeast USRs and have found that localization analysis can be used to further refine and articulate over-representation results.

All k-mers of a given length were scored and sorted based on the over-representation measure $C0/C1$. Surprisingly, most over-represented k-mers have highly non-random distribution patterns. The three most common patterns were chosen for further investigation. The first corresponded to conserved regions in homologous ORF families, the second resulted from certain types of regulatory motifs, and the third resulted from strings containing runs of Ts and As and TAs. Sorting on up/down and on/off expression levels identified many of the same k-mers as the second and third groups. Other distinctive distribution patterns exist, but these provide a reasonable sample of the most common distribution patterns that are selected for using $C0/C1$ on the complete set of yeast's USRs. High-scoring k-mers sometimes result from transposon long terminal repeats, but that source of over-representation was not pursued in this paper.

Most importantly, what emerges from our analysis is a general data-mining methodology for regulatory and other regions in large genomic data sets that is computationally efficient and can flexibly accommodate complementary DNA microarray data. In the case of USRs, the flow chart can be summarized as:

(1) Identify possible interesting k-mers by computing over-representation using several measures, including $C0/C1$, applied to both the coding strand and the coding strand plus its RC.

(2) Analyze the context of these strings using standard alignment and profile methods.

(3) Analyze the spatial distribution of these strings using filters, such as low location variance.

(4) Analyze the structure of these strings using structural scales, such as bendability.

(5) Focus on strings with highly non-random context and/or spatial distribution and/or structural profiles across the USRs containing them.

(6) Strings with highly conserved context and low spatial variance correspond to homologous USRs. The source of homology, such as transposable elements, can be further investigated.

(7) The remaining strings, with highly non-random context/localization/structure can be clustered into different 'patterns' and are likely to play significant roles, including regulatory motifs.

Here we have systematically applied this approach to yeast USRs. No doubt some of the parameters we find in the non-random distribution patterns, such as the '50–200 bp,' are organism-dependent. We are in the process of further corroborating and extending the approach to other genomes, including so far *Escherichia coli*, *Chlamydia*, *Drosophila* and *Arabidopsis* where the methods seem to work as well as for yeast, although the types of k-mers identified are somewhat different. However, the yeast results already show that this approach can contribute to our understanding of the large-scale organization of genomes and the 'lexicon' of regulatory regions.

ACKNOWLEDGEMENTS

The work of P.B. is in part supported by a Laurel Wilkening Faculty Innovation award and a Sun Microsystems

Table 12. Base frequencies and relative entropy over 984 positive-strand occurrences of TATATATA, context of five on each end. Resulting consensus sequence: ATATA TATATATA TATAT

A	50	23	53	28	46	0	100	0	100	0	100	0	100	17	61	18	54	22
T	26	50	20	37	21	100	0	100	0	100	0	100	0	60	17	59	20	54
G	12	11	14	11	20	0	0	0	0	0	0	0	0	8	12	10	14	10
C	10	14	11	22	11	0	0	0	0	0	0	0	0	13	8	11	9	11
RE	0.1	0.1	0.1	0.0	0.1	1.2	1.1	1.2	1.1	1.2	1.1	1.2	1.1	0.2	0.2	0.2	0.1	0.1

Table 13. Sets of ORFs obtained by global alignment of ORFs 1 and 2 (YAL068C and YBR302C) in Table 6 against all 6225 ORFs in the data set

Aligned with 1		Aligned with 2	
YAL068C	500	YBR302C	500
YAR020C	20	YDL248W	466
YBR301W	85	YFL062W	393
YCR104W	374	YGR295C	350
YDR542W	145	YHL048W	482
YEL049W	40	YIR044C	328
YFL020C	20	YJR161C	490
YGL261C	383	YML132W	500
YGR294W	84	YNL336W	488
YHL046C	301		
YIL176C	380		
YIR041W	397		
YJL223C	380		
YKL224C	397		
YLL025W	33		
YLL064C	351		
YLR037C	30		
YLR461W	381		
YMR325W	280		
YNR076W	341		
YOL161C	256		

award at UCI. The work of S.H. and D.K. is partially supported by a grant from the Chao Cancer Foundation.

APPENDIX A: FURTHER ANALYSIS OF HOMOLOGOUS ORF FAMILIES

Once an initial cluster has been identified, a more comprehensive search can be made through the entire data set to find further ORFs with some homology to a representative member of the set. For example, globally aligning the first and second ORFs in Table 7 against all 6225 ORFs in the data set yields the sets in Table 13. This expands the number of ORFs in the two sets from 5 and 8 to 21 and 9.

Given a family of homologous ORFs, it is possible to investigate which parts are most conserved and which parts are variable. Like looking for conserved regions of homologous ORFs across different species, families of homologous ORFs within a single species can be analyzed for conserved regions.

The set of 21 ORFs in Table 13 show a considerable amount of variability in their global alignment scores, but 20 of the 21 contain the 8mer AACAAATA, all at position 10. Likewise, 20 of the 21 contain TATAAATA at position 100. There are numerous conserved strings, and many are in the 0–100 region, suggesting there is something special about this area. This is the basal promoter region, which is in fact quite different from the rest of the USR. For example, unlike the rest of the USR where regulatory motifs may occur on either strand of the DNA, much of the structure of the basal promoter region is asymmetric. It is generally assumed that regulatory motifs occur upstream of this region. The second example above, TATAAATA, is possibly an example of TATA box. Because of its specialized role, it is useful to restrict analysis to just the area between 0 and 100. Various over-represented strings and distinctive distribution patterns are observed in that region which are obscured when the entire 0–500 region is analyzed.

Another use of pairwise alignment scores is to look at the location of homologous ORFs within the complete genome. Specifically, in Figure 8, the location of each ORF that has at least one homologous partner with a local alignment score greater than 350 is shown over the 16 chromosomes of yeast scaled to the same length. Additional members of these homologous ORF sets can be identified by lowering the cutoff from 350, but between 50 and 250, transposable elements are the predominant reason for USR homology. It is apparent that most of the ORFs are in the subtelomeric region of the chromosomes, and most large homologous ORF families are exclusively in these more unstable regions. Furthermore, a detailed inspection shows that most of the homologous ORFs occur in contiguous blocks, indicating that they duplicated as a unit.

APPENDIX B: FURTHER ANALYSIS OF CO-REGULATED ORF FAMILIES

Strings with a symmetric hill-shaped distribution pattern are frequently associated with a rapid decrease in gene expression during oxidative stress (Hampson *et al.*, 2000). For example, for ORFs containing the string GCGATGAGC (Figure 5), three went up and 26 went down, and

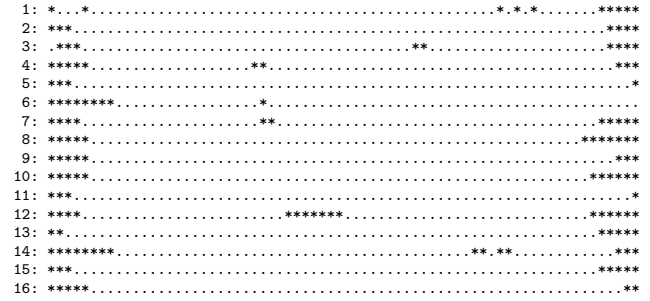


Fig. 8. Location of all 149 ORFs with at least one local USR alignment score greater than 350.

for its RC two went up and 20 went down. This compares to the genome as a whole in which approximately as many went up as down. ORFs were classified as up or down if they showed at least a 1.5-fold change in expression during the first 10 min. 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 between essentially unexpressed ORFs. Using this method 1543 ORFs were classified as changed, 730 up and 813 down. Stricter classification criteria modify the numbers but do not modify the general conclusions.

ORFs that measurably decrease in expression must be initially expressed at a measurable level. Consequently, strings that are correlated with down regulation may also be correlated with significant expression at time zero, i.e. during normal growth conditions. This was investigated by classifying ORFs with an expression less than 30 as off and those greater than 250 as on. Using these somewhat arbitrary cutoffs, 1524 ORFs were on and 1681 were off.

Many strings with symmetric hill-shaped distribution are in fact associated with a high on/off ratio, and conversely, sorting for high on/off values preferentially identifies strings with this distribution pattern. The *k*mer, AAATTTTTC (Figure 4), is a particularly good example of this. It is associated with both high on/off and low up/down values. Conversely, sorting for high on/off or low up/down values preferentially identifies this string and variants of it.

Like sorting on variance or localization in the 50–200 region, sorting on the on/off or up/down ratio is in itself another method of identifying interesting strings, specifically ones that are likely to be regulatory motif instances. For example, the well-known stress element CCCCT has the highest up/down ratio over all 5-mers during oxidative stress and its RC AGGGG has the third highest ratio. While it is possible that a string may be effective on one strand only, sorting on the combined up/down ratio for a string and its RC does appear to favor known motifs. By this measure, the stress element

Table 14. Minimum number of copies of the stress element CCCCT versus up/down ratio

Min copies	Up	Down	Up/down
0	730	813	0.90
1	330	231	1.43
2	112	40	2.80
3	34	5	6.80
4	11	1	11.00
5	4	0	Inf

Table 15. Copies of the IUPAC motif [GT][CA]GATGAG[CAG] versus up/down ratio

Min copies	Up	Down	Up/down
0	730	813	0.90
1	37	175	0.21
2	0	21	0.00
3	0	1	0.00

is first on the list for 5-mers. Further improvement in motif identification is possible if the probability of a string's observed up/down counts is calculated based on the global up/down counts rather than sorting on a simple up/down ratio (Hampson *et al.*, 2000), but motif finding based on up/down or on/off counts is not pursued here. Neither the stress element or its RC have a high $C0/C1$ ratio, showing that although strings with high $C0/C1$ are generally interesting, the converse is not necessarily true.

It was observed that the up/down ratio can show a multiplicity effect: the magnitude of the ratio is correlated with the number of times a string occurs in an ORF. For example, for the stress element, the up/down ratio monotonically increases for the sets of ORFs having a minimum of 0 through 5 copies of the string or its RC (Table 14).

This multiplicity effect provides additional evidence that the string in question is a regulatory motif. Unfortunately, like most long strings, the 9mer GCGATGAGC (Figure 5) never occurs more than once per ORF, so this test is not always applicable. However, by considering all 1-base variations on the original 9mer, it is apparent that several are similarly over-represented, localized, and correlated with reduced expression. These one-base variations can be combined in the IUPAC motif [GT][CA]GATGAG[CAG], which has a much larger number of instances (648) and does show a multiplicity effect (Table 15). As seen in Table 9, there is some conservation of bases in the string's context, and at a minimum, the IUPAC motif should be extended to a 10mer by adding a T on the right end.

REFERENCES

Bailey,T. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn.*, 51–80.

Baldi,P. and Baisnée,P.-F. (2000) Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics*, **16**, 865–889.

Baldi,P. and Brunak,S. (2001) *Bioinformatics: the Machine Learning Approach*, 2nd edn, MIT Press, Cambridge, MA.

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

Brazma,A., Jonassen,I.J., Vilo,J. and Ukkonen,E. (1998) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, **8**, 1202–1215.

Brown,M.P.S., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., M.A.,J. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Brukner,I., Sanchez,R., Suck,D. and Pongor,S. (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J.*, **14**, 1812–1818.

Burge,C., Campbell,A.M. and Karlin,S. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl Acad. Sci. USA*, **89**, 1358–1362.

Bussemaker,H.J., Li,H. and Siggia,E.D. (2000) Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl Acad. Sci. USA*, **97**, 10 096–10 100.

Chen,Q.K., Hertz,G.Z. and Stormo,G.D. (1995) Matrix search 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *CABIOS*, 563–566.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Grove,A., Galeone,A., Mayol,L. and Geiduschek,E.P. (1996) Localized DNA flexibility contributes to target site selection by DNA-bending proteins. *J. Mol. Biol.*, **260**, 120–125.

Hampson,S., Baldi,P., Kibler,D. and Sandmeyer,S. (2000) Analysis of yeast's ORFs upstream regions by parallel processing, microarrays, and computational methods. In *Proceedings of the 2000 Conference on Intelligent Systems for Molecular Biology (ISMB00), La Jolla, CA*. AAAI Press, Menlo Park, CA, pp. 190–201.

Hassan,M.A.E. and Calladine,C.R. (1996) Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J. Mol. Biol.*, **259**, 95–103.

Hegde,P., Qi,R., Abernathy,K., Gay,C., Dharap,S., Gaspard,R., Earle-Hughes,J., Snesrud,E., Lee,N. and Quackenbush,J. (2000) A concise guide to cDNA mircroarray analysis. *Biotechniques*, 548–562.

Hu,Y., Sandmeyer,S., Laughlin,C.M. and Kibler,D. (2000) Combinatorial motif analysis and hypothesis generation on a genomic scale. *Bioinformatics*, **16**, 222–232.

Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.

Iyer,V. and Struhl,K. (1995) Poly (dA : dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.*, **14**, 2570–2579.

Karlin,S. and Mrazek,J. (1997) Compositional differences within and between eukaryotic genomes. *Proc. Natl Acad. Sci. USA*, **94**, 10 227–10 232.

Kibler,D. and Hampson,S. (2001) Learning weight matrices for identifying regulatory elements. In *Proceedings of the 2001 International Conference on Mathematical and Engineering Techniques in Medicine and Biological Science (METMBS-2001)*. pp. 208–214.

Kim,J.M., Vanguri,S., Boeke,J.D., Gabriel,A. and Voytas,D.F. (1998) Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.*, **8**, 464–478.

Lee,M.T., Kuo,F.C., Whitmore,G.A. and Sklar,J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9839.

Long,A.D., Mangalam,H.J., Chan,B.Y., Tolleri,L., Hatfield,G.W. and Baldi,P. (2001) Global gene expression profiling in *Escherichia coli* K12: improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. *J. Biol. Chem.*, **276**, 19 937–19 944.

Nelson,H.C.M., Finch,J.T., Luisi,B.F. and Klug,A. (1987) The structure of an oligo(dA)-oligo(dT) tract and its biological implications. *Nature*, **330**, 221–226.

Olson,W.K., Gorin,A.A., Lu,X., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11 163–11 168.

Parvin,J.D., McCormick,R.J., Sharp,P.A. and Fisher,D.E. (1995) Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor. *Nature*, **373**, 724–727.

Pevzner,P.A. (2000) *Computational Molecular Biology. An Algorithmic Approach*. MIT Press, Cambridge, MA.

Pevzner,P.A. and Sze,S. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. In *Proceedings of the 2000 Conference on Intelligent Systems for Molecular Biology (ISMB00), La Jolla, CA*. AAAI Press, Menlo Park, CA, pp. 269–278.

Sikorski,R.S. and 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.

Starr,D.B., Hoopes,B.C. and Hawley,D.K. (1995) DNA bending is an important component of site-specific recognition by the TATA binding protein. *J. Mol. Biol.*, **250**, 434–446.

van Helden,J., Andre,B. and 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.

van Helden,J., del Olmo,M. and Perez-Ortin,J.E. (2000) Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res.*, **28**, 1000–1010.

Wodicka,L.H. (1997) Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol.*, **15**, 1359–1367.

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

Zhu,Z. and Thiele,D.J. (1996) A specialized nucleosome modulates transcription factor access to a *C. glabrata* metal responsive promoter. *Cell*, **87**, 459–470.