Distribution Patterns of Over-Represented k-mers in Non-Coding Yeast DNA

Steven Hampson Dennis Kibler Pierre Baldi*
Dept. of Information and Computer Science
Institute for Genomics and Bioinformatics
University of California, Irvine
Irvine, CA 92697-3425

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 often associated with the DNA binding sites of transcription factors. We introduce a new measure of over-representation and apply it to the pooled 500bp ORF upstream regions of yeast. More importantly, we investigate the context and spatial distribution of over-represented k-mers in upstream regions.*

Results: *We find that the spatial distributions of most over-represented k-mers are highly non-random. We study the single and double-stranded distribution patterns of these k-mers and relate three especially common patterns to DNA structure, function, and evolution. Specifically, we show that the three most common patterns 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 upstream region; and c) runs of As, Ts, and ATs associated with a broad hill-shaped distribution also in the 50-200 bp upstream region, for which we hypothesize a structural role. Analysis of overrepresentation, 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: *pfbaldi@ics.uci.edu, hampson@ics.uci.edu, kibler@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 compares the frequecy 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 first develop an efficient new measure of over-representation, $C0/C1$, 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.

Somewhat surprisingly, in yeast upstream regions, 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 upstream region; and c) runs of As, Ts, and ATs associated with a broad hill-shaped distribution also in the 50-200 bp upstream region, for which we hypothesize a structural role. This distribution is asymmetric for runs of As and Ts, and trivially symmetric for runs of ATs.

Taken together, the results show that $C0/C1$ over-representation identifies biologically significant k-mers, most of which have distinctive distribution patterns. Analysis of these patterns suggests explanatory mechanisms.

2 Methods

Sequence Data. While over-representation analysis can be applied to any class of genomic DNA, here we focus exclusively on the 500 bp upstream regions (USRs) of each of the 6225 ORFs in the yeast genome

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

taken from the Stanford data base (fttp://genome-ftp.standford.edu/yeast). For convenience, we often use the same label for the ORF and the 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. Some of the methods to be presented have been derived so that they can be applied to raw genomic data. For comparison purposes, 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 (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 Affimetryx Gene Chip microarray technology (Wodicka, 1997). In a typical experiment, the wildtype yeast strain YPH500 (Sikorski & Hieter, 1989) was used 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. GeneChip Expression Analysis v. 3.1 software was used to obtain the average difference values. All experiments were prepared using the polyA mRNA protocol.

Over-Representation. The number of occurrences, $C0$, of each of the 4^k k-mers ($k \leq 9$) can be collected in a single pass through the data set. Only non-overlapping occurrences of each k-mer are counted. Experiments for length $k = 9$ are reported here, but useful results are also obtained with smaller values of k, and would presumably also with larger values. As a point of general comparison, based on the average first-order nucleotide composition of yeast (A≈T≈30%, C≈G≈20%), the expected number of occurrences for a random 9-mer is between approximately 2 and 60. Some strings of interest are well in excess of this range, but many are not.

Over-representation of a k-mer is calculated with respect to some background model. A standard background model is provided by a Markov model of order x. In this case, the over-representation of k-mers can be ranked using the measure $C0/E_x(C0)$, where $E_x(C0)$ is the expected counts based on a model of order x, ($x < k$) (Burge *et al.*, 1992; van Helden *et al.*, 2000; Bussemaker *et al.*, 2000; Baldi & Brunak, 2001).

Here for each k-mer we compute a different ratio, $C0/C1$, where $C1$ is the number of times the string occurs with exactly one mismatch. A random k-mer would be expected to have a $C0/C1$ ratio of approximately $1/3k$ since there are $3 \times k$ different single-mismatch ($M1$) neighbors for each zero mismatch ($M0$) string. An especially large value of $C0/C1$ means that the exact pattern is over-represented compared to the density of patterns in its immediate vicinity. $C0$ 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 $k \leq 9$, this takes only a few seconds on a workstation.

Alternatively, over-representation could be calculated for the non-coding versus coding region, downstream region, or the genome as a whole (Brazma *et al.*, 1998; van Helden *et al.*, 1998; van Helden *et al.*, 2000). All of these measures work to some extent for identifying biologically interesting strings. While the $C0/C1$ measure appears to have some advantages, the main focus of this work is not on the comparison of various measures of over-representation, but on the most common localization patterns of strings derived by sorting on the $C0/C1$ ratio.

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 an over-represented 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 of interest is chosen for further study, we study its context using both local and global alignments of the USRs in which it is found (Durbin *et al.*, 1998). 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. "Highly" is somewhat subjective, but 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

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.

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

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. This local context can be summarized with a consensus string that reflects the most frequent base at each position. Conservation in the window can be assessed by computing at each position the relative entropy (Cover & Thomas, 1991) $RE = \sum_X P_X log(P_X/Q_X)$ between the first-order background distribution Q measured over the entire data set and the observed distribution P over X =A,T,G,C.

3 Results

3.1 Measures of over-representation

Evidence that over-represented strings are biologically significant is provided by the observation that over-represented strings are more apt to be contained in the TRANSFAC data base (Wingender *et al.*, 2001) than strings chosen at random, and under-represented strings are less likely to be found there. This is true whether over-representation is measured by $C0/C1$, $C0/E_1(C0)$, or by $C0$ alone, where $E_1(C0)$ is the expected $C0$ value based on first-order statistics. For k = 7, 8 and 9, the number of strings found in the TRANSFAC database for the 100 strings with the highest and lowest $C0/C1$ ratio, $C0/E_1(C0)$ ratio, $C0$ value, and a set of 100 random strings is shown in Table 1.

A string was counted as matching TRANSFAC if either the string or its RC exactly matched a yeast entry or was a substring of a yeast entry. All three measures support the conclusion that over-represented strings have a higher chance of matching regulatory motifs present in the TRANSFAC database, whereas under-represented strings have a lower chance.

The three measures are correlated but have noticeably different biases. For example, the top 20 9-mers for each measure are shown in Table 2. $C0$, $C0/E_1(C0)$, and $C0/C1$ values are shown for each k-mer. To make the $C0/C1$ ratio more meaningful, it is normalized by dividing by $1/3k$. Sorting on $C0$ yields strings with a strong bias toward high AT/GC ratios, and runs of As and Ts in particular. These strings are presumably biologically significant but, on the whole, are less interesting than over-represented strings with a more balanced AT/GC ratio. This can be addressed by dividing $C0$ by $E_1(C0)$, but the ordering still has a noticeable bias toward strings containing long runs of As and Ts. $C0/C1$ also identifies runs of As and Ts as over-represented (Section 3.4), but none of them are in the top 20. Strings with high $C0$ and $C0/E_1(C0)$ usually have a high $C0/C1$ ratio, so the measures are not independent, but overall, $C0/C1$ appears to yield the most diverse selection of strings.

Using higher-order statistics changes the $C0/E_x(C0)$ ordering, but does not noticeably improve performance over first-order results. Another variation is to sort on the combined counts of a string and its RC. Again, this changes the ordering, emphasizing some types of k-mers over others, but is not necessarily an overall improvement. In any event, the purpose of this paper is to characterize and explore some of the most common patterns of distribution for over-represented strings, not to determine the single best measure of over-representation.

3.2 Overview: Localization and other non-random properties

Strings can be over-represented for a variety of reasons. Consequently, over-represented strings may be associated with a range of different properties depending on the reason for over-representation. Here we focus on k-mers with a distinctive spatial distribution. Three especially common and highly non-random distribution patterns are investigated in detail.

First, k-mers with sharply localized distribution patterns generally identify groups of homologous ORFs. k-mers in the 0-100 base region, corresponding to the basal promoter, appear to be preferentially conserved. When both a k-mer and its RC are tightly localized, divergent ORFs are usually involved. Given one or more members of a homologous ORF family, alignment against the entire data set can be used to find additional members of the family, permitting further analysis of homologous groups. These families of homologous ORFs tend to replicate in blocks near the ends of chromosomes.

Second, k-mers with a broad symmetrical distribution pattern located between 50 and 200 bp appear to identify regulatory motifs. Many of these k-mers are correlated with decreased expression during oxidative stress, or to simply being on during normal growth conditions. The k-mers tend to co-occur in the same USRs, frequently in the context of divergent ORFs. Many of these k-mers appear to be specific instances of more general motifs.

Third, runs of Ts, As, or ATs have a broad [slightly asymmetric for runs of As and Ts/symmetric for runs of ATs] distribution pattern also located between 50 and 200 bp and appear to identify structural elements. These are the most frequent k-mers in the USRs and are fre-

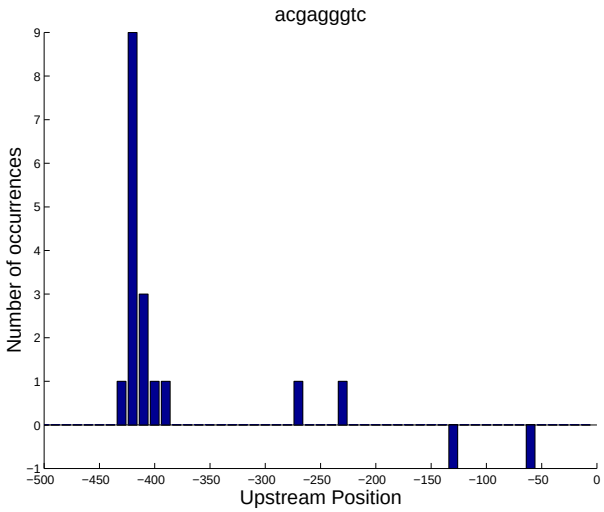
quently correlated with elevated expression during normal growth conditions. Runs of Ts and As are among the stiffest, while alternating Ts and As are among the most flexible sequences of DNA. Their distribution pattern is not due to the distribution of the individual bases.

The number of occurrence of the three categories within the top 20 over-represented 8- and 9-mers is given in Table 3.

3.3 Sharply localized distributions: Homologous ORFs

3.3.1 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 reverse complement GACCCTCGT is shown, along with their $C0$ and $C1$ counts and normalized $C0/C1$ ratio.



9-mer	$C0$	$C1$	$27 \times C0/C1$
ACGAGGGTC	17	87	5.28
GACCCTCGT	2	95	.57

Figure 1: Distribution and counts for ACGAGGGTC and its reverse complement.

A tightly localized distribution on one strand only might result from a group of highly homologous ORFs resulting from duplication events. It is thought that the entire yeast genome underwent an early duplication (Wolfe & Shields, 1997), but that in itself cannot explain anything other than pairs of homologous ORFs. 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 6 of the top 20 $C0/C1$ k-mers in Table 2 are associated with transposon long terminal repeats. The homologous ORF families considered in this paper, however, do not seem to be associated with any transposons contained in the comprehensive list downloadable at http://www.public.iastate.edu/~voytas (Kim *et al.*, 1998). About 7 of the top 20 $C0/C1$ k-mers in Table 2 fall in this category of belonging to large homologous ORF families that do not seem to be associated with transposons.

If the localization pattern does result from a group of homologous ORFs, the shared 9-mer might occur in the context of a larger region of conserved bases. The 5-base context of the string shows that the surrounding area is in fact partially conserved (see Table 4). Better conservation could be achieved by using only those occurrences that produced the localized spike, but for consistency with later examples without sharp spikes, that is not done here. This can be done for both strands of the DNA, but in this example only the top strand is of interest.

All of the context positions are significantly conserved, and a larger window shows an extensive region of base conservation around the string. However, extensive local homology around the string does not necessarily imply overall homology of the USRs. This was investigated using global and local alignment between the entire USRs.

If the set of ORFs containing a highly localized 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, the local and global alignment scores for the USRs of the set of 19 ORFs containing the above string (or its RC) against the first ORF in the set are given in Table 5.

The first ORF YAL068C aligns perfectly with itself yielding global and local scores of 500. It is not expected that an ORF containing the string (like 1) would show any significant alignment with strings containing the RC (6 and 11). Of the 19 ORFs, 13 fall in one homologous family, although the amount of global homology varies over a sizable range. Likewise, the length of the longest highly homologous region varies over a wide range, although this can be adjusted to some extent by adjusting the degree of homology required. 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 local alignment scores of 34 and 23 indicate at least one highly homologous region in those pairs, but does not exclude the possibility of additional shorter regions of equal or better homology or larger regions of lower homology. As expected, the string's location in the 13 homologous ORFs corresponds with the spike in the distribution of the whole set. 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"

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: 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. Numbers in parentheses count the patterns with exact matches in TRANSFAC.

k	8-mer	8-mer + RC	9-mer	9-mer+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

Table 4: Base frequencies and relative entropy (RE) over the 17 positive-strand occurrences of ACGAGGGTC, context of 5 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

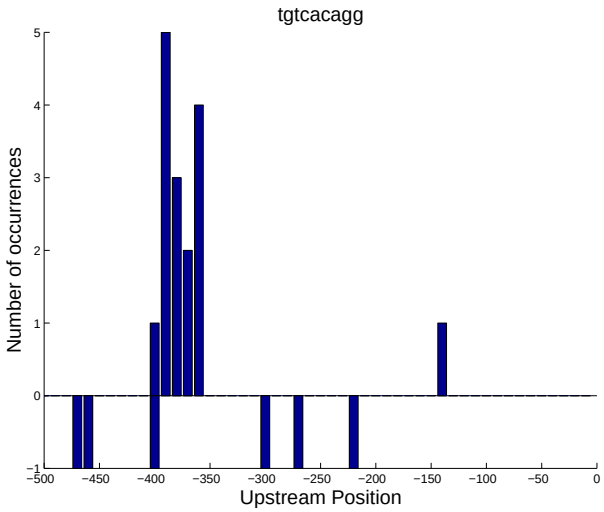
Table 5: Alignment scores of 19 USRs containing the sequence: ACGAGGGTC, or its reverse complement, to the first one of them (YALO68C). G = global, L = local, R = reverse complement.

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

indicating that their coding regions are also homologous. This set of homologous ORFs re-occurs in several of the following examples.

Further exploration of this sort of highly localized distribution pattern was facilitated by specifically selecting for strings with this type of pattern, rather than just visually choosing them from among those selected for high $C0/C1$ ratios. Specifically, the variance in location for each string was calculated and the strings reverse sorted on that value. A minimum of 10 matches was required. This was quite effective in identifying a number of strings with highly localized distribution patterns. Simply replicating a single ORF in the data set will both increase the $C0/C1$ ratio of the strings contained in it and decrease the variance in their location, so 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 6). These two sets correspond to the two closely spaced spikes at about 370. Although highly homologous within clusters, the clusters show little global or local homology between them, beyond the shared 9-mer. In both sets the string's local context is highly conserved, but is completely different between sets (Tables 7 and 8). The fact that a specific 9-mer is localized in approximately the same position in two very different ORF families may be due to random chance, but might indicate functional significance. Further analysis of these homologous families is provided in



9-mer	$C0$	$C1$	$27 \times C0/C1$
TGTCACAGG	16	194	2.23
CCTGTGACA	7	200	.95

Figure 2: Distribution and counts for TGTCACAGG and its reverse complement.

Table 6: 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.

			G1	G2	L1	L2
1	YAL068C		500	-49	500	9
2	YBR302C		-49	500	9	500
3	YCR027C	R	-40	-56	9	8
4	YDL248W		-46	466	14	495
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
10	YGL055W	R	-42	-48	9	8
11	YGL261C		383	-39	500	10
12	YGR295C		-43	350	11	497
13	YHL034C		-47	-34	9	9
14	YHL048W		-51	482	9	500
15	YIR041W		397	-40	498	10
16	YJR161C		-49	490	9	500
17	YKL224C		397	-46	498	10
18	YLL064C		351	-22	474	9
19	YML132W		-49	500	9	500
20	YMR014W	R	-34	-40	8	13
21	YNL336W		-47	488	9	499
22	YPL222W		-46	-45	11	13
23	YPR136C	R	-54	-41	12	8

Table 7: Base frequencies and relative entropy for the 5 occurrences of TGTCACAGG contained in ORFs homologous with YAL068C, context of 5 on each end. Resulting consensus sequence: GGAAA TGTCACAGG CACAG.

A	0	0	100	100	100	0	0	0	0	100	0	100	0	0	0	80	0	80	0
T	0	0	0	0	0	100	0	100	0	0	0	0	0	0	0	0	0	0	0
G	100	100	0	0	0	0	100	0	0	0	0	0	100	100	0	20	0	20	80
C	0	0	0	0	0	0	0	0	100	0	100	0	0	0	100	0	100	0	20
RE	1.7	1.7	1.1	1.1	1.1	1.2	1.7	1.2	1.7	1.1	1.7	1.1	1.7	1.7	1.7	0.8	1.7	0.8	1.2

Table 8: Base frequencies and relative entropy for the 8 occurrences of TGTCACAGG contained in ORFs homologous with YBR302C, context of 5 on each end. Resulting consensus sequence: AGTTT TGTCACAGG AAATC.

A	100	0	0	0	0	0	0	0	0	100	0	100	0	0	75	100	87	0	0
T	0	12	100	100	100	100	0	100	0	0	0	0	0	0	0	0	12	100	0
G	0	87	0	0	0	0	100	0	0	0	0	0	100	100	0	0	0	0	0
C	0	0	0	0	0	0	0	0	100	0	100	0	0	0	25	0	0	0	100
RE	1.1	1.3	1.2	1.2	1.2	1.2	1.7	1.2	1.7	1.1	1.7	1.1	1.7	1.7	0.7	1.1	0.8	1.2	1.7

Appendix A.

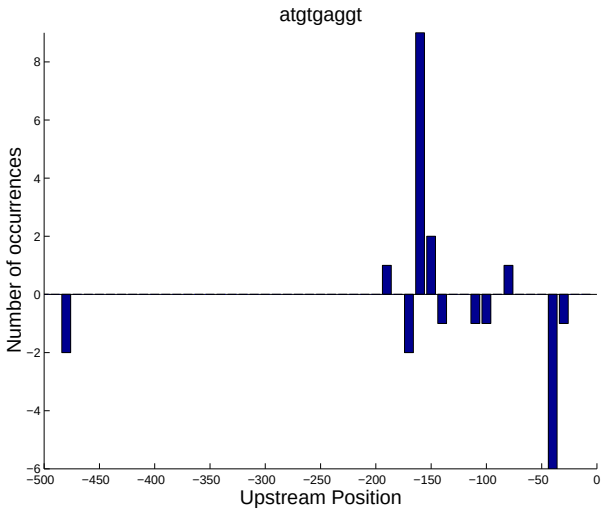
3.3.2 Two-stranded localization: divergent ORFs

Another variation on this type of highly localized distribution pattern 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 the USR of YCR104W (which contains the string on the direct strand) produces a homology set of 9 ORFs, and alignment with the USR of YBL108W (which contains the RC) produces a separate homology set of 6 ORFs (Table 9). The set of 9 ORFs is a subset of the large set of 21 homologous ORFs considered in Appendix A.

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

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

This extended local RC homology suggests that one set might be globally homologous to the RC of the other. However, global alignment does not show any global homology between the two sets if the ORFs are aligned with either the RC of the first or third ORF. Local alignment shows some local homology though. These scores are shown in the final two columns of Table 9, where the ORFs are locally aligned with the RC of the first and third ORFs. Based on this value, it appears that there is in fact considerable local RC homology between the two sets, with the homologous region being roughly



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

Figure 3: Distribution and counts for ATGTGAGGT and its reverse complement.

Table 9: Alignment scores of 27 UTRs containing the sequence ATGTGAGGT, or its reverse complement, to the first and third ones (YBL108W and YCR104W). G = global, L = local, R = reverse complement.

			G1	G3	LR1	LR3
1	YBL108W	R	500	-72	14	208
2	YCR103C	R	392	-63	10	208
3	YCR104W		-72	500	208	12
4	YCRX09C		-82	-66	11	9
5	YDR487C		-84	-91	12	8
6	YGL260W	R	433	-75	14	204
7	YGL261C		-70	397	204	11
8	YHL022C	R	-57	-78	12	11
9	YIL150C	R	-79	-62	8	14
10	YIL174W	R	139	-71	14	310
11	YIL176C		-71	413	204	11
12	YIR040C	R	353	-66	14	205
13	YIR041W		-67	398	205	11
14	YJL223C		-71	413	204	11
15	YJR082C	R	-74	-64	11	12
16	YKL223W	R	347	-67	14	202
17	YKL224C		-63	390	202	11
18	YLL064C		-68	446	208	10
19	YLR461W		-77	401	204	12
20	YMR214W	R	-80	-72	9	9
21	YNR076W		-59	437	205	10
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

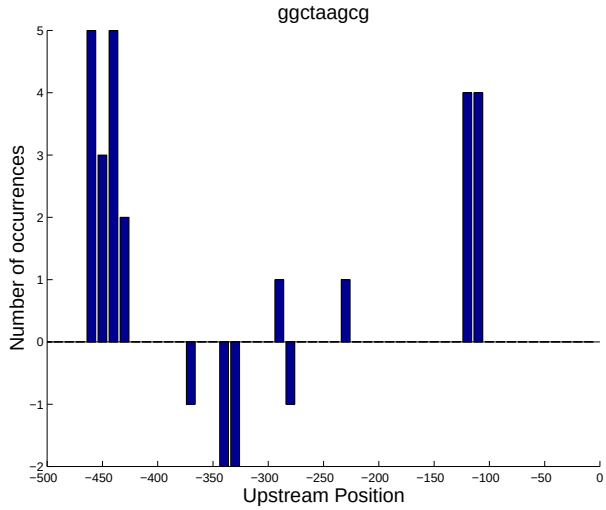
Table 10: ORFs homologous with YBL108W.

YBL108W	500
YCR103C	397
YGL260W	433
YHL045W	410
YIL174W	139
YIR040C	358
YKL223W	352
YMR324C	326

200 bases long. ORF number 10 is an exception with a homologous region of roughly 300 bases.

There may be other mechanisms that can produce this sort of spatial distribution pattern (a string and its RC both tightly localized), but in this case it is associated with a set of divergently transcribed ORFs. Specifically, global alignment of the first ORF with the whole data set produces a slightly larger homology set of 8 ORFs (Table 10). Comparing this expanded set to the expanded homologous family of 21 ORFs in Appendix A, it is apparent that 6 of the 8 ORFs are physically adjacent to ORFs in the large set and on different strands of the DNA (eg. YCR103C and YCR104W). This means their USRs potentially overlap, with the overlapped region of one USR being the RC of the overlapped region of the other. 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 6 divergent pairs all have a distance of approximately 200, resulting in a local RC homology of approximately 200 bases. This also means that the 500-base USR of each ORF extends 300 bases into the coding region of its divergent partner. Regularities within this region might reflect coding constrains as well as regulatory ones. With this arrangement, the sum of the distances to each of the two spikes in the distribution (170 + 40) is equal to the distance between the two ORFs.

Any two adjacent ORFs potentially have such an arrangement if they occur on opposite strands of the DNA, and there are 1343 divergent ORF pairs that overlap to some extent in their 500-base USRs, but the fact that two homologous ORF families would pair this way indicates that the divergent ORFs have frequently replicated together. Note however that there are 8 ORFs in one set, 21 in the other, and only 6 adjacent pairs, so the relationship between sets is not one-to-one. Either the ORFs have also replicated independently, or they replicated as intact pairs but sufficient mutations accumulated so that one member of the pair is no longer recognizable as an ORF. For example, YBL108W (Tables 9 and 10) does not have a divergent partner, but the "empty" region immediately upstream of it is in fact homologous to the divergent partners of the 6 ORFs in Table 10 that do



9-mer	$C0$	$C1$	$27 \times C0/C1$
GGCTAAGCG	25	88	7.67
CGCTTAGCC	6	95	1.71

Figure 4: Distribution and counts for GGCTAAGCG and its reverse complement.

have identified partners. Thus, a pseudogene can be assigned to that region.

Strings are sometimes localized in more than one position (Figure 4). This string was identified using $C0/C1$ since the two widely spaced peaks on the top strand, while highly localized, produce a large variance.

As might be expected, the ORFs fall into three homology classes corresponding to the three peaks in the distribution, two for the string and one for the RC (Table 11). The RC of ORF 3 is both locally and globally homologous with one set of ORFs containing the string but only locally homologous with the other. The two sets containing the string are locally homologous with each other.

The combination of multiple sets with local, global and RC homology in the ORFs sharing the above string is interesting in its own right, but there is one novel feature about the distribution; the string almost always repeats. In Figure 5, the location of the string or its RC is shown in each of the 18 USRs it occurs in. The significance of the repeat is unknown.

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

Table 11: Alignment scores of 18 USRs containing the sequence GGCTAAGCG, or its reverse complement, to the third , fourth, and eight ones (YDR544C, YDR545W, and YGR296W). G = global, L = local, R = reverse complement.

			G3	G4	G8	GR3	L4	LR3
1	YCR060W	R	-62	-41	-39	-57	9	11
2	YDR210W		-59	-69	-46	-62	10	10
3	YDR544C	R	500	-51	-61	-43	8	8
4	YDR545W		-51	500	-47	-46	500	203
5	YER189W		-50	395	-55	-50	478	222
6	YFL064C		-42	411	-46	-43	484	198
7	YFR031C-A		-52	-51	-77	-58	10	10
8	YGR296W		-61	-47	500	385	165	462
9	YIL177C		-41	458	-39	-41	500	198
10	YJL225C		-41	458	-39	-41	500	198
11	YKL107W		-53	-32	-38	-48	10	11
12	YLR467W		-51	500	-47	-46	500	203
13	YML133C		-49	494	-45	-46	499	204
14	YNL338W	R	498	-52	-61	-43	8	8
15	YNL339C		-61	-47	500	385	165	462
16	YPL135W	R	-66	-47	-58	-56	9	10
17	YPL283C		-61	-47	500	385	165	462
18	YPR202W		-62	-55	190	151	162	224

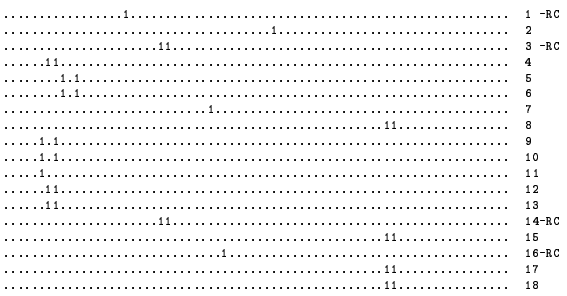


Figure 5: Location of the sequence GGCTAAGCG or its RC in the 18 500bp USRs containing it (Table 11).

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 with overlapping USRs. A number of highly conserved strings are identified. The main point, however, is not to investigate homologous ORFs in depth since there are at least 160 ORFs in this set of subtelomerically-replicated segments. The goal here is to develop a methodology and consider a few examples of some of the most distinctive localization patterns with their possible causes and/or implications.

3.4 Broad Symmetric Distribution: Regulatory Motifs

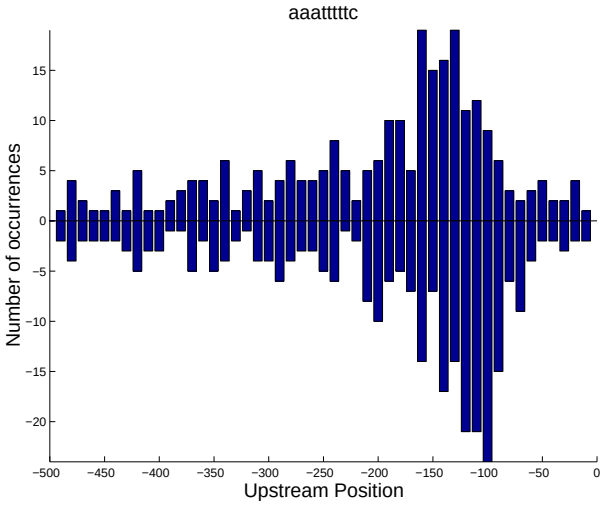
Another frequent spatial distribution pattern for strings with high $C0/C1$ ratios is shown in Figure 6. Including the RC in the calculation of $C0/C1$ helps identify symmetric distributions but it is not a necessity for finding them. In this type of distribution, a string and its RC occur with a broad distribution pattern 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). For 8-mers, this is the most common distribution pattern for strings with high $C0/C1$ ratio. 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. Three examples which appear to be involved in expression regulation will be considered here.

The set of ORFs containing the above string or its RC is too large to include, but shows no global and only limited local alignment homology (12 or 13 bps on the average). The local consensus context shows evidence of weak conservation:

```
S=TTTTTTTTTTAAATATTTGA[-]ATTTTTTTTATAAAATATAA
R=TTTTTTATTTTAATTTTTTA[-]ATTTTTTTTAAAAAATATAA
[-]=AAATTTTTC
```

The similarity between the string and RC context indicates some conservation in bases, but the local context is actually quite variable if the full probability table is considered.

A similar situation arises for another string (Figure 7) which will be considered in more detail. Most of the general conclusions drawn about it are applicable to the previous string. In fact, the two have a strong tendency to co-occur in the same USRs. Both have been previ-



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

Figure 6: Distribution and counts for AAATTTTTC and its reverse complement.

ously identified based on other extraction mechanisms (Hampson *et al.*, 2000; Hughes *et al.*, 2000).

Again 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
```

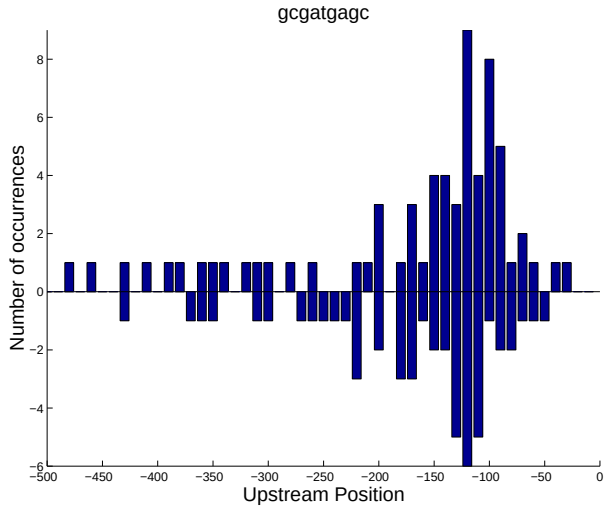
The consensus context indicates some local conservation since the context is similar for the string and its RC. However, it over-states the case since it does not indicate the actual amount of variability. For example, while the 5-base consensus context is identical for the string and its RC in the above example, it is actually quite variable on a case-by-case basis (Table 12).

A random consensus context would generally consist of all As and Ts simply because these are the most common bases, so the fact that the 20-base consensus context consists entirely of As and Ts does not necessarily mean that the area is abnormally AT-rich. The 5-base context, with an average AT frequency of about 71% rather than the expected 60% for a random context is slightly elevated though.

Also, while there is considerable variation at each position in the context, it is interesting to note that the two probability tables are actually quite 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

Table 12: Top half: base frequencies and relative entropy for the 67 positive-strand occurrences of GCGATGAGC, context of 5 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



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

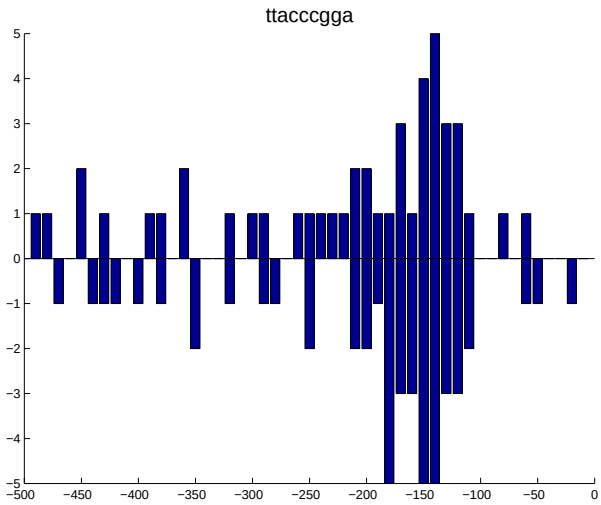
Figure 7: Distribution and counts for GCGATGAGC and its reverse complement.

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 $((2 \times 1343)/6225) \times (290/500) = .25$, or .27 if measured empirically for random 2-mers, 3-mers and 4-mers. By this argument, approximately $(67 + 51)/2 * .27 = 16$ 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 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 this sort of broad distribution pattern are not adequately identified by sorting on location variance, but it is possible to devise specialized measures for this type of distribution. One simple and effective method is to sort on the ratio of counts in the 50-200 region to the total number of counts in the complete USR. Strings with this type of localization pattern that were found using $C0/C1$ almost always had similar distributions for the string and its RC and this was also true for most of the strings selected for the 50-200 region. Consequently, sorting on the combined ratio for the string and its RC is probably justified. A minimum of 40 matches (string + RC) was required. A number of strings can be identified using this specialized metric, almost all having an elevated $C0/C1$ ratio and an unexpectedly high number of divergent pairs.

Many strings identified this way show a strong correlation with expression regulation during oxidative stress, so like 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 all with respect to



9-mer	$C0$	$C1$	$27 \times C0/C1$
TTACCCGGA	45	298	4.64
TCCGGGTAA	50	298	5.15

Figure 8: Distribution and counts for TTACCCGGA and its reverse complement.

oxidative stress, but strings identified by $C0/C1$ will presumably also correlate with other gene expression patterns resulting from other experimental treatments. Such a string is shown in Figure 8. Like the previous string, this is an identified motif (Brazma *et al.*, 1998; Hughes *et al.*, 2000), has 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 decrease in expression during oxidative stress.

To summarize the regulatory motif results, sorting 8-mers and 9-mers on $C0/C1$ identifies a number of strings with a broad, approximately symmetric, localization pattern in the 50-200 bp upstream of the transcription start point. 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.

3.5 Broad Asymmetric Distribution: Strings of Ts and As

Ts and As are over-represented in the USR with frequencies of approximately A≈T≈30% and C≈G≈20%. Because of this bias, strings with high AT/GC ratio have elevated C0 and $C0/C1$ values. 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 downstream 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 & 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 ($3 < k \leq 10$), and runs of Ts containing a single C and runs of As containing a single G occupy many of the other top spots. For example, of the top 20 9-mers when sorted on $C0$ (Table 2), all Ts and all As are first and second, while Ts with a C and As with a G fill 14 of the remaining positions. Alternating TAs fill 2 more. 18 of the 20 form RC pairs even though the strings were sorted on their individual $C0$ values.

Consecutive runs of Ts and As have a distinct, somewhat asymmetric distribution pattern (Figure 9). Runs of Ts containing a C, and runs of As containing a G show similar distribution patterns.

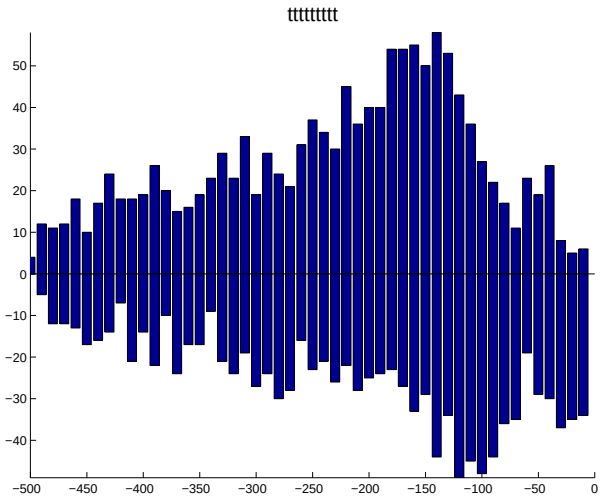
Consecutive runs of Ts and As are highly over-represented: for 9-mers, 1320 and 1250 versus an expected 60 based on the first-order frequency of Ts and As. They do not occur in the context of large-scale homology or in an unexpected number of divergent ORF pairs. They are not strongly correlated with 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 on/off value.

The preference for interrupting a run of Ts with a C can be seen in the local context of the string TTTTCTTTT (Table 13). There is only partial base conservation, but the enhanced probability of Ts and Cs is still apparent.

The tendency for runs of Ts and As is reflected in the probability of seeing a T or A based on what precedes

Table 13: Base frequencies and relative entropy for the 629 positive-strand occurrences of TTTTCTTTT, context of 5 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



9-mer	$C0$	$C1$	$27 \times C0/C1$
TTTTTTTTT	1320	12397	2.87
AAAAAAAAA	1250	12060	2.80

Figure 9: Distribution and counts for TTTTTTTTT and its reverse complement.

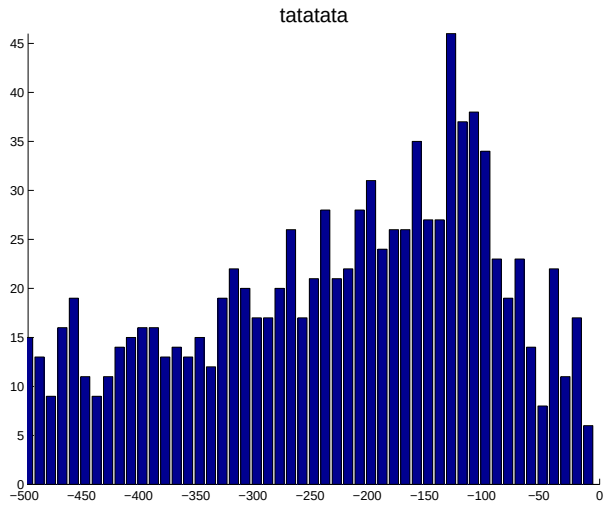
it (Table 14). The more Ts that precede a location, the more likely it is to be another T. In addition, a run of Ts is more likely to be followed by a C than a G. Similar results are obtained for runs of As. Cs and Gs show a different pattern for the length of their runs.

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 long 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 $k \leq 4$, 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 in more likely. A similar situation occurs with runs of As.

Alternating Ts and As show a similar distribution

Table 14: Probability of extending a preceding run of As, Cs, Gs, and Ts.

min # of prec. Ts	prob A	prob T	prob G	prob C
0	.3166	.3128	.1834	.1872
1	.2602	.3662	.1858	.1882
2	.2107	.4047	.1823	.2031
3	.1933	.4394	.1604	.2083
4	.1695	.4753	.1443	.2124
5	.1442	.5360	.1199	.2016
6	.1168	.6082	.0995	.1769
7	.0965	.6453	.0852	.1748
min # of prec. As	prob A	prob T	prob G	prob C
0	.3166	.3128	.1834	.1872
1	.3643	.2806	.1858	.1684
2	.4000	.2448	.1970	.1568
3	.4320	.2284	.1927	.1435
4	.4646	.2135	.1914	.1265
5	.5239	.1983	.1721	.1029
6	.6040	.1659	.1501	.0776
7	.6448	.1492	.1385	.0655
min # of prec. Gs	prob A	prob T	prob G	prob C
0	.3166	.3128	.1834	.1872
1	.3167	.2784	.1946	.2106
2	.3186	.2770	.1869	.2181
3	.3141	.2809	.1860	.2195
4	.3274	.2504	.1929	.2300
5	.3281	.2142	.2373	.2197
6	.3493	.1603	.3378	.1489
7	.3057	.1189	.4755	.1019
min # of prec. Cs	prob A	prob T	prob G	prob C
0	.3166	.3128	.1834	.1872
1	.3304	.3116	.1639	.1942
2	.3312	.3144	.1613	.1930
3	.3108	.3385	.1598	.1922
4	.3069	.3294	.1559	.2083
5	.2866	.3097	.1479	.2566
6	.2302	.3611	.1128	.2979
7	.1822	.3036	.0607	.4555



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

Figure 10: Distribution and counts for TATATATA (equal to its reverse complement).

(Figure 10) perfectly symmetric, however, since it is reverse-complement invariant–hence only the string is shown. 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.

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

The 5-base context is TA-rich and shows the preference for T-A alternation (Table 15). 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.

These sequences all have remarkable structural properties. Runs of A's (or T's) are the stiffest as can be ascertained using a number of dinucleotide or trinucleotide structural scales (Baldi *et al.*, 1999; Baldi & Baisnée, 2000), ranging from DNAse I bendability (Brukner *et al.*, 1995), to propeller twist angle (Hassan & 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. A number of promoters in yeast contain homopolymeric dA:dT elements. Such homopolymeric tracts are known from X-ray crystallography to be straight and rigid (Nelson *et al.*, 1987). 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 & Struhl, 1995; Zhu & 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. The triplet ATA/TAT, characteristic of the TATA box, is highly flexible according to the bendability scale, 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 & Baisnée, 2000) and are likely to be associated with flexible stretches of DNA, the role of which remains to be determined.

4 Discussion

Over-representation is an effective method for identifying biologically important k-mers in raw genomic DNA sequences. In addition, localization analysis can be used to refine the results obtained from over-representation analysis alone.

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. Three 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 is not pursued in this paper.

The approach used here was motivated in part by the method described in (van Helden *et al.*, 1998). In that 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 the whole genome 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. This can be calculated for all k-mers and sorted.

This approach works well for highly conserved motifs (those with little variability and thus only a few different motif instances) but degrades as variability increases, since the over-representation of each individual

Table 15: Base frequencies and relative entropy over 984 positive-strand occurrences of TATATATA, context of 5 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

motif instance decreases as the number of different motif instances increases. In order to extend the power of the approach, over-representation of the $M1$ and $M2$ set, corresponding respectively to one or two base pair chances, for each $M0$ string was also computed (Hampson *et al.*, 2000). Based on artificial data, this was effective in identifying motifs and motif instances that could not be identified using over-representation of the M0 strings only.

This established the utility of using $M0$, $M1$ (and sometimes $M2$) statistics, and the general idea of using over-representation to identify potential motif instances. Consequently, when considering the genome as a whole, the idea of "single mismatch over-representation" based on the $C0/C1$ ratio was a simple and at least plausible criteria on which to order strings. Its main attraction is that it is fast and can be applied directly to the data set without the need for clustering or expression data. Run time is linear with the size of the data set, so large sets can be accommodated. Of course global analysis of the entire data set is likely to preferentially identify the most general features, so separate analysis of individual regions or subsets is also productive when meaningful groups are available. Similarities and differences with other organisms is also of interest.

Other measures of over-representation are possible, such as sorting on $C0$ alone, the $C0/E_x(C0)$ ratio, or the probability of seeing C0 occurrences given the $C1$ or $E_x(C0)$ counts. These other measures sort the strings in different orders, identifying different k-mers of interest, and possibly finding other types of distribution patterns. However, the $C0/C1$ ratio was sufficiently effective in identifying biologically interesting k-mers that alternative over-representation measures were not extensively explored.

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. As a specific point of interest, it was observed that strings with a high $C0/C1$ ratio generally have distinctive spatial distribution patterns. Some of these patterns could be explained based on known properties of yeast USRs, and some could not. Based on our initial investigations, several of these patterns proved to be productive points of departure for exploring other properties of the USR, and it seems likely that other strings with these distribution patterns or other distribution patterns that were not considered here will be equally interesting.

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 summarize as:

1. Identify possible interesting k-mers, for instance by computing over-representation using $C0/C1$, or some other measure, applied to both the coding strand and the coding strand plus its reverse complement.
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 applied systematically 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. Thus we are in the process of further corroborating and extending the approach to other genomes. But 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.

Table 16: Sets of ORFs obtained by global alignment of ORFs 1 and 2 (YAL068C and YBR302C) 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		

5 Appendix A: Further analysis of homologous ORF families

Once an initial cluster has been identified, a more comprehensive (and slower) search can be made through the entire data set to find further ORFs that have some homology to a representative member of the set. For example, globally aligning the first and second ORFs in Table 6 against all 6225 ORFs in the data set yields the sets in Table 16. This expands the number of ORFs in the two sets from 5 and 8 to 21 and 9.

It takes several minutes to align one ORF against the entire data set, so while it is possible to extract a homologous ORF cluster by aligning one or more cluster members against the complete data set, it is generally impractical to discover clusters by aligning all ORFs against all other ORFs, which in this case would take at least a week. Given sufficient resources, such exhaustive alignment can be done (Hampson *et al.*, 2000) but similar results can be achieved in a few minutes by using $C0/C1$ or variance to identify conserved strings and candidate clusters which are then fleshed out by exhaustive alignment using selected cluster members.

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 16 show a considerable amount of variability in their global alignment scores when aligned with the first on the list, indicating a wide range of homology, but 20 of the 21 contain the 8-mer AACAAATA, all at position 10. Likewise, 20 of the 21 contain TATAAATA at position 100. In both cases the RC does not occur at all. There are numerous conserved strings in the set, and many of the most highly conserved strings 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.

The best representation of ORF families based on USR homology would probably be a hierarchy reflecting the duplication points and subsequent divergence of the homologous ORFs. For example, based on pairwise global alignment scores, a bottom-up clustering of one set of homologous USRs produces the binary tree in Figure 11, which provides a possible phylogeny for the set of ORFs. The tree is constructed by sequentially merging sets that produce the largest average pairwise global alignment score over all USRs in the merged set. The sets are initialized to the individual USRs.

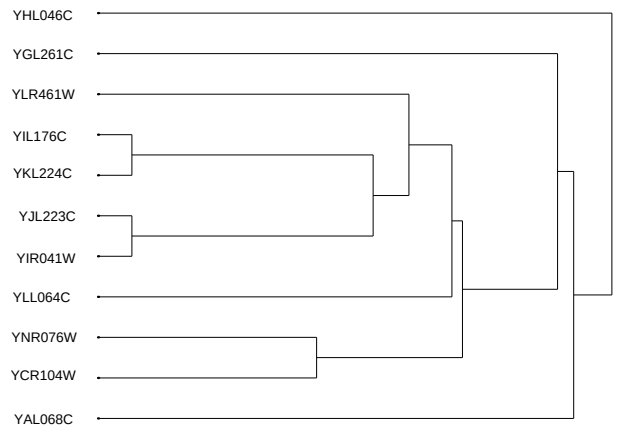


Figure 11: Binary tree resulting from hierarchical clustering of homologous USRs.

Another use of pairwise alignment scores is to look at the location of homologous ORFs within the complete genome. Specifically, in Figure 12, the location of each ORF that has at least one homologous partner with a

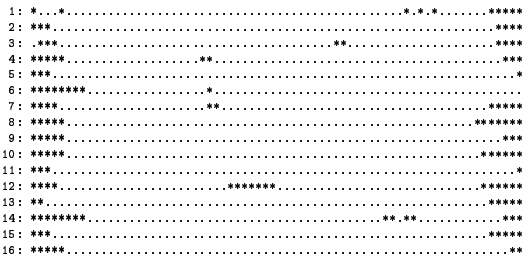


Figure 12: Location of all 149 ORFs with at least one local USR alignment score greater than 350.

local alignment score greater than 350 is shown over the 16 chromosomes of yeast. Additional members of these homologous ORF sets could be identified by lowering the cutoff from 350, but between 50 and 250, transposable elements are the predominant reason for USR homology. All chromosomes are scaled to the same length. 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 that region. The biological significance of this localization is not known. Furthermore, a detailed inspection shows that most of the homologous ORFs occur in small contiguous blocks of 4 or 5 ORFs, indicating that they duplicated as a unit.

6 Appendix B: Further analysis of co-regulated ORF families

Strings with this sort of distribution pattern are frequently associated with rapid changes in gene expression during oxidative stress, almost always with a decrease in expression (Hampson *et al.*, 2000). For example, for ORFs containing the string GCGATGAGC (Figure 7), 3 went up and 26 went down, and for its RC 2 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 ten minutes. 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 changed the absolute numbers but did not change the general conclusions.

Strings with this spatial distribution pattern were often associated with decreased expression, and the converse was also generally true. If the strings were sorted on their up/down ratio, those with high ratios showed little localization while those with low ratios frequently showed the above pattern.

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 this type of 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 6), 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 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 17).

This multiplicity effect provides additional evidence that the string in question is a regulatory motif. Unfortunately, like most long strings, the 9-mer GCGATGAGC (Figure 7) never occurs more than once per ORF, so this test is not always applicable. However, by considering all 1-base variations on the original 9-mer (that is, its $M1$ set), it is apparent that several are similarly overrepresented, localized and correlated with reduced expression. These one-base variations can be combined in the IUPAC motif [GT][CA]GATGAG[CAG], which has

Table 17: Minimum number of copies of the stress element CCCCT versus up/down ratio.

min copies	up	down	up/down
0	730	813	.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 18: Copies of the IUPAC motif [GT][CA]GATGAG[CAG] versus up/down ratio.

min copies	up	down	up/down
0	730	813	.90
1	37	175	.21
2	0	21	.00
3	0	1	.00

a much larger number of instances (648) and does show a multiplicity effect (Table 18). As seen in Table 12, there is some conservation of bases in the string's context, and at a minimum, the IUPAC motif should be extended to a 10-mer by adding a T on the right end.

This highlights a drawback of identifying exact k-mers as regulatory motifs: in many cases the actual motif may be degenerate. For example, the above IUPAC motif matches 12 different 9-mers, and if the motif if statistically interesting, all of its subsets, supersets and shifted versions will also be to a lesser extent. The relationship among these exact k-mers is not always apparent since they may be intermixed with instances of other degenerate motifs.

A more general criticism is that individual k-mers might be functionally inactive and statistically unremarkable, but important in combination. Such strings would not be found, but there are sufficient strings that are individually interesting that the issue can be deferred.

The standard approach to motif identification is to first identify families of co-regulated genes. The most common approach is to cluster expression data (DeRisi *et al.*, 1997),(Eisen *et al.*, 1998), (Brown *et al.*, 2000), (Hu *et al.*, 2000). 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; Baldi & Long, 2001). Alternatively, a literature search was used in (van Helden *et al.*, 1998) to produce 10 families of co-regulated yeast genes that could be used for analysis see also (Hughes *et al.*, 2000)).

Determining the optimum motif description based on limited data is also a difficult process with many different approaches. In (van Helden *et al.*, 1998), binding sites are represented as exact k-mers, and a statistical measure of over-representation is used to find them. 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 & Elkan, 1995; Chen *et al.*, 1995; Brown *et al.*, 2000; Hu *et al.*, 2000). 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 & Sze, 2000; Pevzner, 2000) and references therein). We have developed a hill-climbing algorithm to optimize IUPAC motifs for over-representation ((Hampson *et al.*, 2000)), but the process of motif optimization is not pursued here.

Acknowledgements

The work of PB is in part supported by a Laurel Wilkening Faculty Innovation award and a Sun Microsystems award at UCI. The work of SH and DK is partially supported by a grant from the Chao Cancer Foundation.

References

Bailey, T. & Elkan, C. (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 51–80.

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

Baldi, P. & Brunak, S. (2001). Bioinformatics: the machine learning approach. MIT Press, Cambridge, MA. Second edition.

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

Baldi, P. & Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

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

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

Brukner, I., Sanchez, R., Suck, D. & 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. & Karlin, S. (1992). Over- and under-representation of short oligonucleotides in dna sequences. *PNAS*, **89**, 1358–1362.

Bussemaker, H. J., Li, H. & Siggia, E. D. (2000). Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *PNAS*, **97**, 10096–10100.

Chen, Q. K., Hertz, G. Z. & 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.

Cover, T. M. & Thomas, J. A. (1991). Elements of Information Theory. John Wiley & Sons, Inc.

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

Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci USA*, **95**, 14863–14868.

Grove, A., Galeone, A., Mayol, L. & 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. & 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. & 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. & Quackenbush, J. (2000). A concise guide to cDNA mircroarray analysis. *Biotechniques*, 548–562.

Hu, Y., Sandmeyer, S., Laughlin, C. M. & 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. & 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. & 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. & Mrazek, J. (1997). Compositional differences within and between eukaryotic genomes. *PNAS*, **94**, 10227–10232.

Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A. & 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. & Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *PNAS*, **97**, 9834–9839.

Long, A. D., Mangalam, H. J., Chan, B. Y., Tolleri, L., Hatfield, G. W. & 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. *Journal of Biological Chemistry*, **276**, 19937–19944.

Nelson, H. C. M., Finch, J. T., Luisi, B. F. & 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. & Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA*, **95**, 11163–11168.

Parvin, J. D., McCormick, R. J., Sharp, P. A. & 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. The MIT Press, Cambridge, MA.

Pevzner, P. A. & 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. & 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. & 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. & 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. & Perez-Ortin, J. E. (2000). Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res.*, **28**, 1000–1010.

Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., Thiele, S. & Urbach, S. (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–284.

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

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

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