A Statistical Method for Finding Transcription Factor Binding Sites *

Saurabh Sinha and **Martin Tompa**
Department of Computer Science and Engineering
Box 352350
University of Washington
Seattle, WA 98195-2350 U.S.A.
206-543-9263
fax: 206-543-8331
{saurabh,tompa}@cs.washington.edu

Abstract

Understanding the mechanisms that determine the regulation of gene expression is an important and challenging problem. A fundamental subproblem is to identify DNA-binding sites for unknown regulatory factors, given a collection of genes believed to be coregulated, and given the noncoding DNA sequences near those genes. We present an enumerative statistical method for identifying good candidates for such transcription factor binding sites. Unlike local search techniques such as Expectation Maximization and Gibbs samplers that may not reach a global optimum, the method proposed here is guaranteed to produce the motifs with greatest z-scores. We discuss the results of experiments in which this algorithm was used to locate candidate binding sites in several well studied pathways of *S. cerevisiae*, as well as gene clusters from some of the hybridization microarray experiments.

Keywords: sequence analysis, motif, transcription factor, binding site, promoter, spacer, z-score.

1. Transcription Factor Binding Sites

1.1. Identifying Eukaryotic Regulatory Sequences

One of the major challenges facing biologists is to understand the mechanisms for the regulation of gene expression. In particular, for any given biochemical pathway, there are often complex interactions among its set of genes and their products.

There have been a number of recent studies that used DNA microarrays to identify the sets of genes involved in certain pathways of the yeast *S. cerevisiae* (DeRisi *et al.* (1997), Chu *et al.* (1998), Spellman *et al.* (1998)). These studies divided the set of genes into subsets whose expression patterns suggest that they may be coregulated.

The next step in unraveling the regulatory interactions is to identify common binding sites in the regulatory regions of these coregulated genes and, from these binding sites, identify the regulatory factor that binds there (Chu *et al.* (1998), Roth *et al.* (1998), Spellman *et al.* (1998), Tavazoie *et al.* (1999)). It is precisely this problem of identifying unknown transcription factor binding sites that we address.

The analysis of noncoding regions in eukaryotic genomes in order to identify regulatory sequences is a difficult problem, and one that is by no means well understood. There are several reasons for this difficulty:

1. The regulatory sequences may be located quite far from the corresponding coding region, either upstream or downstream or in the introns.
2. The regulatory sequences need not be in the same orientation as the coding sequence or each other.
3. There may be multiple binding sites for a single factor in a single gene's regulatory region.
4. There can be great variability in the binding sites of a single factor, and the nature of the allowable variations is not well understood.

In *S. cerevisiae*, the first of these problems is not severe: nearly all transcription factor binding sites are believed to lie within 800 bp upstream of the translation start site (Zhu and Zhang (1999)). The three remaining confounding problems are, however, present.

1.2. Previous Methods for Finding Regulatory Motifs

A number of algorithms to find general motifs have been proposed previously. (See, for example, Bai-

* This material is based upon work supported in part by the National Science Foundation and DARPA under grant DBI-9601046, and in part by the National Science Foundation under grant DBI-9974498.

ley and Elkan (1995), Fraenkel *et al.* (1995), Galas *et al.* (1985), Hertz and Stormo (1999), Lawrence *et al.* (1993), Lawrence and Reilly (1990), Rigoutsos and Floratos (1998), Rocke and Tompa (1998), and Staden (1989).) Many of these algorithms are designed to find longer and more general motifs than are required for identifying transcription factor binding sites. The types of general motif used in the cited references include weight matrices, alignments, and gapped alignments. The price paid for this generality is that many of the cited algorithms are not guaranteed to find globally optimal solutions, since they employ some form of local search, such as Gibbs sampling, expectation maximization, and greedy algorithms, that may end in a local optimum. There have been a few studies that have applied these local search techniques specifically to the problem of identifying transcription factor binding sites in *S. cerevisiae* (Chu *et al.* (1998), Roth *et al.* (1998), Spellman *et al.* (1998), Tavazoie *et al.* (1999)), with some success.

The number of well conserved bases in the collection of binding sites of a single *S. cerevisiae* transcription factor is typically six to ten (Wingender *et al.* (1996), Zhu and Zhang (1999)). This number is small enough that, for this particular problem, one need not rely on such general local search heuristics. Instead, one can afford to use enumerative methods that guarantee global optimality. This is the approach taken by the current paper, whose method is most closely allied to that of van Helden *et al.* (1998) and Tompa (1999).

Van Helden *et al.* (1998) used an enumerative statistical method to tackle the same problem of finding transcription factor binding sites in *S. cerevisiae*. Their method proved reasonably successful at finding short, contiguous transcription factor binding sites. However, their method suffers from some drawbacks that we rectify:

1. They consider only exact matches, disallowing variations in the binding site instances of a given transcription factor.
2. Their motifs do not include "spacers", which precludes their algorithm from finding such well known binding sites as that of Gal4p, whose consensus is `CGGNNNNNNNNNNNCCG` (Wingender *et al.* 1996; Zhu & Zhang 1999).
3. In their statistical model, they assume that occurrences of a motif at distinct sequence positions are probabilistically independent, whereas in reality overlapping occurrences (in both orientations) have rather complex dependencies (Nicodème *et al.* (1999)).
4. Their measure of statistical significance of a motif s is based on the frequency of occurrence of s over all regulatory regions of the genome. This is problematic for those motifs that appear rarely, because there may be insufficient data to support reliable statistics. (See Salzberg *et al.* (1998) for a discussion.) The more standard Markov chain model that we employ can be based on the frequencies of shorter (and hence more frequent) oligonucleotides.

Brāzma *et al.* (1998) employed a similar technique for identifying binding sites. They did allow their motifs to contain up to three occurrences of the `N` character.

Tompa (1999) used an enumerative method similar to that of van Helden *et al.* (1998), but for finding ribosome binding sites in prokaryotic genomes. We adopt some of that work's statistical considerations here, in particular, the use of a Markov chain to model the background genomic distribution, the use of z-score as the measure of statistical significance, and attention to the autocorrelation of overlapping motif instances. However, Tompa's algorithm also suffers some shortcomings for the present application:

1. Tompa's algorithm also did not allow for spacers in the motifs, since they seemed irrelevant in the prokaryotic ribosome binding site problem.
2. The allowable variability among binding site instances that proved sufficient for the prokaryotic ribosome binding site problem, namely zero or one substitution from some consensus sequence, proved insufficient in the present application.
3. The possibility of multiple binding sites for a single factor in a single gene's regulatory region does not arise in the prokaryotic ribosome binding site problem. This complicates the motif autocorrelation computation.

2. Motifs and Their Significance

2.1. Variability Among Binding Site Instances

The first question that must be addressed is "What constitutes a motif?" for the application of transcription factor binding sites in *S. cerevisiae*. An inspection of transcription factor databases (such as TRANSFAC (Wingender *et al.* 1996) or SCPD (Zhu & Zhang 1999)), or of the relevant literature (particularly Jones *et al.* (1992), which is rich in examples, and also Blaiseau *et al.* (1997), Mai and Breeden (1997), McInerny *et al.* (1997), Nurrish and Treisman (1995), Oshima *et al.* (1996), and Wemmie *et al.* (1994)) reveals that there is significant variation among the binding sites of any single transcription factor, so that it is overly rigid to insist on exact matches among motif instances. Moreover, the nature of the variability itself varies

from factor to factor, so that the "correct" motif model is far from clear.

Certain trends that must be incorporated in the motif model do, however, emerge:

1. Like the Gal4p binding site consensus mentioned in Section 1.2, many of the motifs have spacers varying in length from 1 to 11 bp. The spacer usually occurs at the middle of the motif, often because the factors bind as dimers.
2. The number of well conserved bases (not including spacers, of course) is usually in the range 6–10.
3. When there is variation in a conserved motif position, it is often a transition (that is, the substitution of a purine for a purine, or a pyrimidine for a pyrimidine) rather than a transversion. This is because of the similarity in nucleotide size necessary to fit the transcription factor's fixed DNA-binding domain. Somewhat less often, the variation in a given position may be between a pair of complementary bases. Other positional variations are rarer.
4. Insertions and deletions among binding sites are uncommon, again because of the fixed structure of the factor's DNA-binding domain.

Based on these observations, a motif for our application will be a string over the alphabet {A, C, G, T, R, Y, S, W, N}, with 0-11 consecutive N's at the center, and a limited number of R's (purine), Y's (pyrimidine), S's (strong), and W's (weak). We choose such a consensus model rather than (say) a weight matrix in order to be able to enumerate motifs. Note that there is little need to allow further variation in motif instances, since the variation is already incorporated in the motif's allowance of R,Y,S,W,N. An examination of 50 binding site consensi included in SCPD (Zhu & Zhang 1999) revealed that the number of consensi that exactly fit this characterization is 31 (62%). About 10 more fit the characterization if very slight differences from the exact consensus are tolerated.

2.2. Measure of Statistical Significance

Given some set of (presumably coregulated) *S. cerevisiae* genes, the input to our problem is the corresponding set of upstream sequences, each having length 800 bp and having its 3′ end at the gene's translation start site.

A good measure for comparing motifs must take into account both the absolute number of occurrences of the motif in the input sequences, and the background genomic distribution. (See Tompa (1999) for a detailed discussion.) For each motif s, let N_s be the number of occurrences of s in the input sequences, allowing an arbitrary number of occurrences per upstream sequence. A reasonable measure of s as a motif, then, would reflect how unlikely it would be to have N_s occurrences, if the sequences were instead drawn at random according to the background distribution.

More specifically, let X be a set of random DNA sequences of the same number and lengths as the input sequences, but generated by a Markov chain of order m, whose transition probabilities are determined by the $(m+1)$-mer frequencies in the full complement of 6000+ upstream regions (each of length 800 bp) in *S. cerevisiae*. (In our experiments, we chose $m = 3$ in order for the background model to include the TATA, AAAA, and TTTT sequences that are ubiquitous throughout the genome's upstream regions (van Helden, André, & Collado-Vides 1998).) Let the random variable X_s be the number of occurrences of the motif s in X, and let $E(X_s)$ and $\sigma(X_s)$ be its mean and standard deviation, respectively. Then the *z-score* associated with s is

$$z_s = \frac{N_s - E(X_s)}{\sigma(X_s)}. \qquad (1)$$

The measure z_s is the number of standard deviations by which the observed value N_s exceeds its expectation, and is sometimes called the "normal deviate" or "deviation in standard units". See Leung *et al.* (1996) for a detailed discussion of this statistic. The z-score z_s obeys, in the asymptotic limit, a normal distribution. This is known to be the case when X is a singleton set: see Nicodème *et al.* (1999, Theorem 2). The result extends to an arbitrary finite set X (with equal sized regions) by a Central Limit Theorem due to Lindeberg (Feller 1993, Section X.1, Formula 1.4). The measure z_s is normalized to have mean 0 and standard deviation 1, making it suitable for comparing different motifs s.

What remains to discuss, then, is how to compute the mean $E(X_s)$ and standard deviation $\sigma(X_s)$. The former is straightforward but the latter, because of the possibility of overlap of a motif with itself (in either orientation), is not. Fortunately, this problem of pattern autocorrelation has been well studied, beginning with its introduction by Guibas and Odlyzko (1981). (See the excellent overview by Nicodème *et al.* (1999).) In particular, a method for computing the standard deviation $\sigma(X_s)$ that is more efficient than using the general recurrence formulae of Nicodème *et al.* (1999) was presented by Kleffe and Borodovsky (1992) for first-order Markov chains and the case in which the motif s is a single string. We have generalized their formulae to our case, in which s represents a finite set of strings. (See also Régnier (1998).) Note that, in this case, one must take into account all possibilities

of one string in the set overlapping with any other. Our extension allows higher order Markov chains, spacers to be handled at no extra run-time cost, and the possibility of a motif occurring in either orientation, none of which were relevant considerations for Kleffe and Borodovsky. All these changes taken together result in a substantial modification of the formulae of Kleffe and Borodovsky: see the Appendix for details.

2.3. Algorithm Summary

The complete algorithm is summarized as follows. The inputs to the algorithm are

1. a set of upstream sequences,
2. the number of nonspacer characters in the motifs to be enumerated, and
3. the transition matrix for an order m Markov chain constructed from the full complement of upstream sequences of *S. cerevisiae*.

The algorithm first makes a pass over the input sequences, tabulating the number N_s of occurrences of each motif s in either orientation. For each motif s for which $N_s > 0$, it then uses the method described in the Appendix to compute $E(X_s)$ and $\sigma(X_s)$, and uses Equation (1) to compute the z-score z_s. It outputs the motifs sorted by z-score.

For a single motif s, the running time to compute z_s is $O(c^2k^2)$, where k is the number of nonspacer characters in s, and c is the number of possible instantiations of `R`, `Y`, `S`, and `W` symbols in s. Because the number of motifs is exponential in k, we can afford this enumerative method only for modest values of k. Note, however, that the dependence on genome size is linear, so that the method scales very well to large genomes.

Moreover, the $O(c^2k^2)$ time z-score computation does not need to be computed for most of the motifs. A very significant reduction in running time is achieved by the following optimization: We note that the dominant part of a motif's z-score computation is the variance calculation. We also note that z_s can be bounded by the expression

$$z_s \leq \frac{N_s - E(X_s)}{\sqrt{E(X_s) - E(X_s)^2}} \qquad (2)$$

since $\sigma(X_s)^2 \geq E(X_s) - E(X_s)^2$. (See Equation (4) in the Appendix.) Hence, before computing $\sigma(X_s)$, we compute $E(X_s)$ and use Inequality (2) to examine if it may be worthwhile to go into the variance computation. (We compare this expression to the lowest z-score among the top ranking motifs discovered so far.) If not, the variance computation for s is aborted, and the next motif is examined.

A similar bounding technique is used to optimize the variance computation itself. Noting that the dominant part of the variance computation is computing the *overlap term* $\sum_{i=1}^{|CW|} E(X_i^{(CW)})$ (see Equation (5) in the Appendix), which is nonnegative, we compute the remaining terms of the variance first, and compute the overlap term only if there is a possibility of getting a high enough z-score. (The overlap term contributes to the denominator of the expression in Equation (1), so the z-score is maximized when the overlap term is 0.) Our experiments showed that these two optimizations reduce the running time of the algorithm drastically.

3. Experimental Results

3.1. Known Regulons

We implemented and ran the program described in Section 2.3 on seventeen well studied coregulated sets of genes in *S. cerevisiae*. For each of these seventeen sets of upstream sequences, there is a known transcription factor with a known binding site consensus, so that the success of the experiments can be assessed.

In all but two of these experiments, our algorithm succeeded in determining the known consensus, in the following sense: In nine of them, the known consensus was one of the three highest scoring motifs; and in six others a very similar looking motif was in the top three. Tables $1-8$ give examples of some of these successes. In each table, the known consensus is given in the caption, and its instances in the program's output are italicized. As can be seen, often the known consensus and its close relatives are prominent in the five highest-scoring motifs. (We chose the number of nonspacer characters in order to make the comparison with the known consensus easier. Choosing a slightly different number produces similar results.)

Note the unusually high z-scores in many of these tables; one would not expect scores so many standard deviations above the mean in random data. To verify this assertion, for each family we ran the program on several independent sets of simulated data generated by the 3rd order Markov chain described for the random variable X in Section 2.2. For each such simulated input, we computed the maximum z-score, and then the mean of these maxima. We call this the *mean max z-score* for the family, and include it in the caption of each table. Note the disparity between this mean max z-score and the actual z-scores of the top motifs in most of the tables.

In the remaining two experiments (ACE2 and ADR1, both being families with very few genes in

s	N_s	z_s
TCANNNNNNACG	27	9.67
TCRNNNNNNACG	34	9.36
YCANNNNNNACG	34	8.58
TCANNNNNNWCG	37	8.39
YCANNNNNNWCG	52	8.31

Table 1: Five highest scoring motifs for the 19-gene family ABF1, whose known consensus is TCANNNNNNACG (Zhu & Zhang 1999). Mean max z-score on simulated data : 6.37

s	N_s	z_s
CGGNNNNNNNNNNCCG	28	32.72
CGGNNNNNNNNNNNSCG	31	28.72
CGGNNNNNNNNNNNCSG	28	26.03
CGGNNNNNNNNNNNCCS	28	25.52
CGGNNNNNNNNNNNYCG	29	25.13

Table 2: Five highest scoring motifs for the 6-gene family GAL4, whose known consensus is CGGNNNNNNNNNNCCG (Zhu & Zhang 1999). Mean max z-score on simulated data : 6.84

s	N_s	z_s
ACGCGT	26	19.24
ACGCGW	35	17.63
ACGCGY	30	15.74
ACGSGT	30	14.39
CGCGTY	31	14.38

Table 3: Five highest scoring motifs for the 12-gene family MCB, whose known consensus is ACGCGT [Chris Roberts, personal communication]. Mean max z-score on simulated data: 6.48

s	N_s	z_s
TCCGYGGA	14	38.02
TCCGCGGA	8	34.16
TCCRYGGA	20	33.53
TCCGYGGR	15	32.02
TCCRCGGR	15	31.81

Table 4: Five highest scoring motifs for the 7-gene family PDR3, whose known consensus is TCCGYGGA (Zhu & Zhang 1999). Mean max z-score on simulated data: 14.32

s	N_s	z_s
CACGTGGG	3	16.75
CCGCNNNNNNNNNTGCC	3	16.66
CACGTGSG	4	16.56
CCGGNNNNCGGC	2	16.36
CACGTGGR	5	16.34

Table 5: Five highest scoring motifs for the 5-gene family PHO, whose known consensus is GCACGTGGG or GCACGTTTT (Oshima, Nobuo, & Harashima 1996). Mean max z-score on simulated data: 16.0

s	N_s	z_s
CACGAAA	10	15.92
CCGNNNNCGGA	4	15.11
CRCGAAA	12	14.95
CWCGAAA	12	13.37
CGTNNNNNNCGCA	4	13.21

Table 6: Five highest scoring motifs for the 3-gene family SCB (or SWI), whose known consensus is CNCGAAA (Zhu & Zhang 1999). Mean max z-score on simulated data: 10.98

s	N_s	z_s
TGAAACA	15	9.17
AACNNNNNNNWRAC	22	8.91
TGAAACR	18	8.61
TRAAACA	23	8.59
TRAAWCA	30	8.35

Table 7: Five highest scoring motifs for the 9-gene family STE12, whose known consensus is TGAAACA [Chris Roberts, personal communication]. Mean max z-score on simulated data: 8.9

s	N_s	z_s
TCACGTG	19	23.63
TCRCGTG	20	20.33
TCACGYG	20	20.07
ATANAYAT	62	19.28
ATANNNAYAT	57	18.87

Table 8: Five highest scoring motifs for the 11-gene family MET, whose known consensus is TCACGTG or AAAACTGTGG (van Helden, André, & Collado-Vides 1998). Mean max z-score on simulated data: 8.26

s	N_s	z_s
ACGCGT	104	34.88
ACGCGW	149	34.02
ACGCGY	121	28.57
RCGCGW	172	26.77
CGCGTY	119	24.63

Table 9: Five highest scoring motifs for the 57-gene cluster CLN2 (Spellman *et al.* 1998). The cluster is regulated by MCB, SCB and the binding site could be WCGCGW (MCB) or CNCGAAA (SCB) (Zhu & Zhang 1999). See also Table 3.

them), the known consensus was in the top twenty reported motifs.

3.2. Coexpressed Gene Clusters

We also ran our program on eight of the coexpressed gene clusters discovered by Spellman *et al.* (1998) and Tavazoie *et al.* (1999).

Tables 9 – 12 summarize the results from the best four of these experiments (three from (Spellman *et al.* 1998) and one from (Tavazoie *et al.* 1999)). Again, the top five motifs in each family have very high z-scores and match the binding site consensus of the transcription factor believed to regulate the family. In three out of four of these experiments, the authors found a very similar motif. The fourth experiment is on the Y' cluster from Spellman *et al.* (1998), whose regulation is not well understood, and for which the authors reported no striking motif. Table 10 does reveal some very conspicuous and high scoring motifs. These turn out to be part of a repeated 168- to 173-mer, which occurs in close variations in 18 of the 31 upstream regions.

4. Future Work

The results of our approach have been most promising. There are several issues and aspects that warrant further research:

- The current motif characterization is still limited. In some true binding sites, spacers may not be centered, or there may be more than one run of spacers. We do not handle such motifs yet.
- We are investigating how much of the work done in the enumerative loop of the algorithm can be moved to the preprocessing step, before the coregulated gene sequences are input. We believe the program can be made much faster this way.
- The accuracy of the results could be improved by filtering out well known repeats from the upstream regions of the genes before running our tool on them.

s	N_s	z_s
GACGNNNNNNGGAC	23	56.33
CTGCNNNNNGCAG	36	55.85
GCAGNNNCTGC	36	55.67
CAGANTCTG	36	51.93
CAGANNCTGC	36	50.29

Table 10: Five highest scoring motifs for the 31-gene cluster Y' (Spellman *et al.* 1998). The regulator and binding site for the cluster are unknown.

s	N_s	z_s
R*ARCCAGC*	23	14.82
ARCCAGCA	17	13.75
ARCCAGCR	20	12.94
RRCCAGCA	20	12.33
ARAANAARA	138	12.23

Table 11: Five highest scoring motifs for the 27-gene cluster SIC1 (Spellman *et al.* 1998). The cluster is regulated by Swi5p/Ace2p and the binding site is believed to be RRCCAGCR.

s	N_s	z_s
ACGCGW	51	10.19
ACGCGT	32	9.77
CGCGTY	49	9.02
ACGCGW	175	29.87
ACGCGT	114	28.77
RCGCGW	207	23.48
ACGCGT	116	29.33
ACGCGW	164	27.68
ACGCGY	140	24.51

Table 12: Three highest scoring motifs for each of three subsets of the the 186-gene cluster 2, which is involved in replication and DNA synthesis (Tavazoie *et al.* 1999). The three subsets mimic the authors' cross-validation experiment. The cluster is regulated by MCB, SCB and the binding site could be WCGCGW (MCB) or CNCGAAA (SCB) (Zhu & Zhang 1999). See also Tables 3 and 9.

- More experiments need to be done to determine a good threshold for significant z-scores. This threshold should depend on the number of non-spacer characters as well as the size of the input sequences.
- We are experimenting with more gene families for which the binding site is not yet known, including families from other eukaryotic genomes.
- In some of the experiments some motifs with very high significance were discovered, but they are not documented as binding sites. These motifs need closer examination.

Acknowledgments

We thank Linda Breeden and Chris Roberts for sharing their insights on transcription factor binding sites, and Rimli Sengupta for numerous thorough and extremely helpful discussions.

A. Appendix

This section describes how we compute $E(X_s)$ and $\sigma(X_s)$ for a given motif s, when X is a single region of length n. The motif s is assumed to be a string of length l over the alphabet {A,C,G,T,R,Y,S,W,N}. For simplicity, the Markov model assumed here is of 1st order; in Section A.4 the changes necessary to accommodate higher orders are described.

The motif s is first converted into a set W of strings, which contains strings of length l over the alphabet {A,C,G,T,N}, by replacing the R's, Y's, S's, and W's by all possible combinations of the appropriate bases. Then for each string in W, its reverse complement is also added to W. (As a result, W may be a multiset.) Notice that motif s occurs at a given position in X (on either strand) if and only if some string in the set W occurs at that position.

A.1. Number of Occurrences

X_s is defined as the sum of the number of occurrences (in X) of each member of W. (Overlapping instances are counted as separate.) Since palindromes occur twice in W, we are effectively counting two for each occurrence of every such palindrome. The reason for this is that an occurrence of a palindrome on one strand accounts for two occurrences of the motif when both strands are considered.

We denote members of W by W_i, and $|W|$ by T (counting duplicate elements as distinct).

Define X_{ij}, for $i \in \{1, 2, \ldots, T\}$ and $j \in \{1, 2, \ldots, n-l+1\}$, as a 0/1 indicator variable for the occurrence of W_i at position j, i.e.,

$$X_{ij} = \begin{cases} 1, & \text{if } W_i \text{ occurs at position } j \text{ of } X \\ 0, & \text{otherwise} \end{cases}.$$

Also,

$$X_{sj} \stackrel{def}{=} \sum_{i=1}^{T} X_{ij},$$

$$X_i \stackrel{def}{=} \sum_{j=1}^{n-l+1} X_{ij},$$

$$X_s \stackrel{def}{=} \sum_{i=1}^{T} X_i.$$

This definition of X_s is consistent with the definition in Section 2.2. X_s counts the total number of occurrences of the motif s in X, taking both strands into account, and considering the special case of palindromes also.

(Note that $X_s = \sum_{i=1}^{T} X_i = \sum_{i=1}^{T}\sum_{j=1}^{n-l+1} X_{ij} = \sum_{j=1}^{n-l+1}\sum_{i=1}^{T} X_{ij} = \sum_{j=1}^{n-l+1} X_{sj}$.)

A.2. Expectation

By definitions and the linearity of expectation, we have

$$E(X_s) = \sum_{i=1}^{T} E(X_i),$$

$$E(X_i) = \sum_{j=1}^{n-l+1} E(X_{ij}),$$

$$E(X_{ij}) = \Pr(X_{ij} = 1) = p_j(a_{i1})p_*(W_i),$$

where $p_j(c)$ is the probability of occurrence of base c at position j, a_{im} is the mth character of string W_i and $p_*(W_i)$ is the probability of W_i starting at any position, given that a_{i1} occurs at that position.

Assuming p_j to be a constant independent of j, we can denote p_j by p and rewrite the formula above as $E(X_{ij}) = p(a_{i1})p_*(W_i)$, from which we get

$$E(X_i) = (n - l + 1)p(a_{i1})p_*(W_i). \qquad (3)$$

The assumption that p_j is independent of j is discussed and justified by Kleffe and Borodovsky (1992). The vector p is the so-called *stationary distribution* of the Markov chain.

We compute $p_*(W_i)$ by following the Markov chain for $l-1$ steps starting with a_{i1}. In following the Markov chain, we have to "skip over" any spacers by using higher powers of the transition matrix (which can be precomputed for efficiency).

A.3. Variance

The variance of X_s is, by definition,

$$\sigma(X_s)^2 = E(X_s^2) - E(X_s)^2,$$

where

$$\begin{aligned} E(X_s^2) &= E((\sum_{i=1}^{n-l+1} X_{si})^2) \\ &= \sum_{j=1}^{n-l+1}\sum_{k=1}^{n-l+1} E(X_{sj}X_{sk}) \\ &= \sum_{i=1}^{n-l+1} E(X_{si}^2) \\ &\quad +2\sum_{j=1}^{n-l+1}\sum_{k=j+1}^{n-l+1} E(X_{sj}X_{sk}). \end{aligned}$$

Let B be the first summation in this expression, and $2C$ be the remaining terms. We first consider the term C.

$$\begin{aligned} C &= \sum_{j=1}^{n-l+1}\sum_{k=j+1}^{n-l+1} E(X_{sj}X_{sk}) \\ &= \sum_{j=1}^{n-l+1}\sum_{k=j+1}^{j+l-1} E(X_{sj}X_{sk}) \\ &\quad + \sum_{j=1}^{n-2l+1}\sum_{k=j+l}^{n-l+1} E(X_{sj}X_{sk}) \\ &= \sum_{j=1}^{n-l+1}\sum_{k=1}^{l-1}\sum_{i_1=1}^{T}\sum_{i_2=1}^{T} E(X_{i_1,j}X_{i_2,j+k}) + A, \end{aligned}$$

where

$$A = \sum_{j=1}^{n-2l+1}\sum_{k=j+l}^{n-l+1} E(X_{sj}X_{sk}).$$

Now let CW be the set of all overlapping concatenations of pairs of strings from W. That is,

$$CW = \{xyz \mid W_{i_1} = xy \text{ and } W_{i_2} = yz, \text{ for some } i_1, i_2, \text{ and nonempty } x, y, z\}.$$

We denote members of CW by CW_i. Like W, CW can potentially be a multiset.

Also, define

$$X_{ij}^{(CW)} = \begin{cases} 1, & \text{if } CW_i \text{ occurs at position } j \\ 0, & \text{otherwise} \end{cases}.$$

Notice that there is a one-to-one correspondence between

$$\{(k, i_1, i_2) \mid X_{i_1,j}X_{i_2,j+k} = 1 \text{ and } 0 < k < l\}$$

and

$$\{i \mid X_{ij}^{(CW)} = 1\},$$

for any j. Note also that the event $X_{i_1,j}X_{i_2,j+k} = 1$ is identical to the event $X_{ij}^{(CW)} = 1$, for the corresponding i.

Therefore,

$$\sum_{k=1}^{l-1}\sum_{i_1=1}^{T}\sum_{i_2=1}^{T} E(X_{i_1,j}X_{i_2,j+k}) = \sum_{i=1}^{|CW|} E(X_{ij}^{(CW)})$$

where $|CW|$ denotes the cardinality of CW.

We can thus write

$$\begin{aligned} C &= \sum_{j=1}^{n-l+1}\sum_{i=1}^{|CW|} E(X_{ij}^{(CW)}) + A \\ &= \sum_{i}\sum_{j=1}^{n-|CW_i|+1} E(X_{ij}^{(CW)}) + A, \end{aligned}$$

where $|CW_i|$ denotes the length of the string CW_i. Let $X_i^{(CW)} = \sum_{j=1}^{n-|CW_i|+1} X_{ij}^{(CW)}$. Then we have

$$\sum_{j=1}^{n-|CW_i|+1} E(X_{ij}^{(CW)}) = E(X_i^{(CW)}),$$

which can be computed just as any $E(X_i)$ is computed. (See Equation (3).) Let $p^k(c_2|c_1)$ denote the probability of finding the character c_2 k steps (of the Markov chain) after c_1. Defining $q = n - 2l + 2$, we can then write C as

$$C = \sum_{i=1}^{|CW|} E(X_i^{(CW)}) + A,$$

where

$$\begin{aligned} A &= \sum_{j=1}^{n-2l+1}\sum_{k=j+l}^{n-l+1} E(X_{sj}X_{sk}) \\ &= \sum_{j=1}^{q-1}\sum_{k=1}^{q-j}\sum_{i_1=1}^{T}\sum_{i_2=1}^{T} E(X_{i_1,j}X_{i_2,k+j+l-1}) \\ &= \sum_{j}\sum_{k}\sum_{i_1}\sum_{i_2} p_j(a_{i_1,1})p_*(W_{i_1})p^k(a_{i_2,1}|a_{i_1,l})p_*(W_{i_2}) \\ &= \sum_{i_1}\sum_{i_2} p_*(W_{i_1})p_*(W_{i_2})S_{i_1 i_2}, \end{aligned}$$

where

$$S_{i_1 i_2} = \sum_{j=1}^{q-1}\sum_{k=1}^{q-j} p^k(a_{i_2,1}|a_{i_1,l})p_j(a_{i_1,1}).$$

By imitating the computation shown in the proof of Theorem 1 in Kleffe and Borodovsky (1992),

and making appropriate approximation (replacing a certain power series sum by its asymptotic limit, as explained in Kleffe and Borodovsky (1992)), we finally get

$$S_{i_1 i_2} = p(a_{i_1,1})\{\frac{q(q+1)}{2}p(a_{i_2,1}) - (q-1)\mathbf{e}'_{\mathbf{a_{i_1 1}}}\mathbf{QPe}_{\mathbf{a_{i_2 1}}} - \mathbf{e}'_{\mathbf{a_{i_1 1}}}\mathbf{QP^2Qe}_{\mathbf{a_{i_2 1}}}\}.$$

Here, $\mathbf{e}_{\mathbf{a_{i_1 1}}}$ and $\mathbf{e}_{\mathbf{a_{i_2 1}}}$ are elementary unit vectors of length 4 that have a 1 in the position corresponding to the last character of W_{i_1} and the first character of W_{i_2} respectively, $\mathbf{P}$ is the 4×4 transition probability matrix of the Markov chain and $\mathbf{Q}$ is $(\mathbf{P} - \mathbf{I} - \mathbf{1p}')^{-1}$, as defined in Kleffe and Borodovsky (1992). (The notations $\mathbf{e}'$ and $\mathbf{p}'$ denote the transpose of those vectors, and the vector $\mathbf{1}$ is the column vector with all 1s.)

Now consider the term B defined earlier.

$$\begin{aligned} B &= \sum_{j=1}^{n-l+1} E(X_{sj}^2) \\ &= \sum_j E(\sum_{k=1}^{T} X_{kj}^2 + \sum_{q=1}^{T}\sum_{r\neq q} X_{qj}X_{rj}) \\ &= \sum_j \sum_k E(X_{kj}^2) + \sum_j \sum_q E(\sum_{r\neq q} X_{qj}X_{rj}) \\ &= \sum_j \sum_k E(X_{kj}) + \sum_j \sum_q E(\sum_{r\neq q} X_{qj}X_{rj}). \end{aligned}$$

To simplify the second term, when $r \neq q$,

- $X_{qj}X_{rj} = 1$ if W_q and W_r are strings that can both be instantiated at position j, and
- $X_{qj}X_{rj} = 0$ otherwise.

The simplest case in which $X_{qj}X_{rj} = 1$ is when W_q and W_r are two copies of the same palindrome, and it occurs starting at position j. An example of the more general case is the motif $s =$ `AASANNSTT`. Two of its four instantiations in W would be

$$W_1 = \texttt{AACANNCTT}$$

and

$$W_2 = \texttt{AAGANNGTT}.$$

The reverse complement of W_1 would then also be added to W, namely

$$W_5 = \texttt{AAGNNTGTT}.$$

Now it is possible for both W_2 and W_5 to be instantiated starting at position j, even though W_2 and W_5 are not identical. We will say that W_2 and W_5 can be *superimposed*.

If W_q and W_r can be superimposed, with $r \neq q$, then they cannot both be instances of the motif s, or both be reverse complements of instances of s. Hence, for every q, there is at most one $r \neq q$ such that W_q and W_r can be superimposed. Let PAL be the set of indices q such that W_q can be superimposed with W_r, for some $r \neq q$.

Rather than checking all pairs in W to find which can be superimposed, it is more efficient to identify such pairs directly from the motif s. This is easily done by reading s from both ends at once. For each pair of superimposable strings W_q and W_r so identified, it is also easy to determine the most general common instantiation P_q of both W_q and W_r. For the example strings W_2 and W_5 above, $P_2 = P_5 =$ `AAGANTGTT`. For $q \in PAL$, let

$$Y_{qj} = \begin{cases} 1, & \text{if } P_q \text{ occurs at position } j \\ 0, & \text{otherwise} \end{cases}.$$

and

$$Y_q = \sum_{j=1}^{n-l+1} Y_{qj}.$$

Then we can write

$$\begin{aligned} B &= \sum_{j=1}^{n-l+1}\sum_{k=1}^{T} E(X_{kj}) + \sum_{j=1}^{n-l+1}\sum_{q\in PAL} E(Y_{qj}) \\ &= E(X_s) + \sum_{q\in PAL} E(Y_q). \end{aligned}$$

$E(Y_q)$ can be computed just as any $E(X_i)$ is computed, using Equation (3).

In summary, the variance of X_s can be obtained from the following set of formulae:

$$\sigma(X_s)^2 = E(X_s^2) - E(X_s)^2,$$

$$E(X_s^2) = B + 2C,$$

$$B = E(X_s) + \sum_{q\in PAL} E(Y_q), \quad (4)$$

$$C = \sum_{i=1}^{|CW|} E(X_i^{(CW)}) + A, \quad (5)$$

$$A = \sum_{i_1=1}^{T}\sum_{i_2=1}^{T} p_*(W_{i_1})p_*(W_{i_2})S_{i_1 i_2},$$

$$S_{i_1 i_2} = p(a_{i_1,1})\{\frac{q(q+1)}{2}p(a_{i_2,1}) - (q-1)\mathbf{e}'_{\mathbf{a_{i_1 1}}}\mathbf{QPe}_{\mathbf{a_{i_2 1}}} - \mathbf{e}'_{\mathbf{a_{i_1 1}}}\mathbf{QP^2Qe}_{\mathbf{a_{i_2 1}}}\}.$$

A.4. Higher Order Markov Model

This section outlines how to extend the calculations above to handle higher order Markov chains. In the

expectation calculations of Section A.2, $p_j(c)$ now denotes the probability of occurrence of the m-mer c at position j, a_{i1} in Equation (3) is now the first m-mer of W_i, and $p_*(W_i)$ is computed, as before, by following the Markov chain, starting with a_{i1}. The transition matrix $\mathbf{P}$ is now a $4^m \times 4^m$ matrix, where the rows and columns are indexed by the possible m-mers, and $\mathbf{P}_{ij}$ is the probability that the m-mer j starts at position $t+1$, given that the m-mer i starts at position t. Thus, each row in $\mathbf{P}$ has at most 4 nonzero entries.

The variance calculations given in Section A.3 remain the same, except for $S_{i_1 i_2}$, which depends on m. For the case $m = 3$ used in our experiments, it is given by

$$\begin{aligned} S_{i_1 i_2} &= p(a_{i_1,1})\{\frac{q(q+1)}{2} p(a_{i_2,1}) - \\ &\quad (q-1)\mathbf{e}'_{\mathbf{a_{i_1 1}}}\mathbf{Q}\mathbf{P}^{\mathbf{3}}\mathbf{e}_{\mathbf{a_{i_2 1}}} - \\ &\quad \mathbf{e}'_{\mathbf{a_{i_1 1}}}\mathbf{Q}\mathbf{P}^{\mathbf{2}}\mathbf{Q}\mathbf{P}^{\mathbf{2}}\mathbf{e}_{\mathbf{a_{i_2 1}}}\}. \end{aligned}$$

References

Bailey, T. L., and Elkan, C. 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* 21(1–2):51–80.

Blaiseau, P.-L.; Isnard, A.-D.; Surdin-Kerjan, Y.; and Thomas, D. 1997. Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Molecular and Cellular Biology* 17(7):3640–3648.

Brāzma, A.; Jonassen, I.; Vilo, J.; and Ukkonen, E. 1998. Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Research* 15:1202–1215.

Chu, S.; DeRisi, J.; Eisen, M.; Mulholland, J.; Botstein, D.; Brown, P. O.; and Herskowitz, I. 1998. The transcriptional program of sporulation in budding yeast. *Science* 282:699–705.

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.

Feller, W. 1993. *An Introduction To Probability Theory And Its Applications*, volume 1. Wiley Easter Limited, third edition.

Fraenkel, Y. M.; Mandel, Y.; Friedberg, D.; and Margalit, H. 1995. Identification of common motifs in unaligned DNA sequences: application to *Escherichia coli* Lrp regulon. *Computer Applications in the Biosciences* 11(4):379–387.

Galas, D. J.; Eggert, M.; and Waterman, M. S. 1985. Rigorous pattern-recognition methods for DNA sequences: Analysis of promoter sequences from Escherichia coli. *Journal of Molecular Biology* 186(1):117–128.

Guibas, L. J., and Odlyzko, A. M. 1981. String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory, Series A* 30:183–208.

Hertz, G. Z., and Stormo, G. D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15(7/8):563–577.

Jones, E. W.; Pringle, J. R.; and Broach, J. R., eds. 1992. *The Molecular and Cellular Biology of the Yeast* Saccharomyces*: Gene Expression*. Cold Spring Harbor, NY: Cold Spring Harbor Press.

Kleffe, J., and Borodovsky, M. 1992. First and second moment of counts of words in random texts generated by Markov chains. *Computer Applications in the Biosciences* 8(5):433–441.

Lawrence, C. E., and Reilly, A. A. 1990. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function, and Genetics* 7:41–51.

Lawrence, C. E.; Altschul, S. F.; Boguski, M. S.; Liu, J. S.; Neuwald, A. F.; and Wootton, J. C. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262:208–214.

Leung, M.-Y.; Marsh, G. M.; and Speed, T. P. 1996. Over- and underrepresentation of short DNA words in herpesvirus genomes. *Journal of Computational Biology* 3(3):345–360.

Mai, B., and Breeden, L. 1997. Xbp1, a stress-induced transcriptional repressor of the *Saccharomyces cerevisiae* Swi4/Mbp1 family. *Molecular and Cellular Biology* 17(11):6491–6501.

McInerny, C. J.; Partridge, J. F.; Mikesell, G. E.; Creemer, D. P.; and Breeden, L. L. 1997. A novel Mcm1-dependent element in the SWI4, CLN3, CDC6, and CDC47 promoters activates M/G_1-specific transcription. *Genes & Development* 11:1277–1288.

Nicodème, P.; Salvy, B.; and Flajolet, P. 1999. Motif statistics. Technical Report RR-3606, INRIA Rocquencourt.

Nurrish, S. J., and Treisman, R. 1995. DNA binding specificity determinants in MADS-box transcription factors. *Molecular and Cellular Biology* 15:4076–4085.

Oshima, Y.; Nobuo, O.; and Harashima, S. 1996. Regulation of phosphatase synthesis in *Saccharomyces cerevisiae* – a review. *Gene* 179:171–177.

Régnier, M. 1998. A unified approach to word statistics. In *RECOMB98: Proceedings of the Sec-*

ond Annual International Conference on Computational Molecular Biology, 207–213.

Rigoutsos, I., and Floratos, A. 1998. Motif discovery without alignment or enumeration. In *RECOMB98: Proceedings of the Second Annual International Conference on Computational Molecular Biology*, 221–227.

Rocke, E., and Tompa, M. 1998. An algorithm for finding novel gapped motifs in DNA sequences. In *RECOMB98: Proceedings of the Second Annual International Conference on Computational Molecular Biology*, 228–233.

Roth, F. P.; Hughes, J. D.; Estep, P. W.; and Church, G. M. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology* 16:939–945.

Salzberg, S. L.; Delcher, A. L.; Kasif, S.; and Owen, W. 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research* 26(2):544–548.

Spellman, P. T.; Sherlock, G.; Zhang, M. Q.; Iyer, V. R.; Anders, K.; Eisen, M. B.; Brown, P. O.; Botstein, D.; and Futcher, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* 9:3273–3297.

Staden, R. 1989. Methods for discovering novel motifs in nucleic acid sequences. *Computer Applications in the Biosciences* 5(4):293–298.

Tavazoie, S.; Hughes, J. D.; Campbell, M. J.; Cho, R. J.; and Church, G. M. 1999. Systematic determination of genetic network architecture. *Nature Genetics* 22:281–285.

Tompa, M. 1999. An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, 262–271. Heidelberg, Germany: AAAI Press.

van Helden, J.; André, B.; and Collado-Vides, J. 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology* 281(5):827–842.

Wemmie, J. A.; Szczypka, M. S.; Thiele, D. J.; and Moye-Rowley, W. S. 1994. Cadmium tolerance mediated by the yeast AP-1 protein requires the presence of an ATP-binding cassette transporter-encoding gene, *YCS1*. *Journal of Biological Chemistry* 269(51):32592–32597.

Wingender, E.; Dietze, P.; Karas, H.; and Knüppel, R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Research* 24(1):238–241. `http://transfac.gbf-braunschweig.de/TRANSFAC/`

Zhu, J., and Zhang, M. Q. 1999. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* 15(7/8):563–577. `http://cgsigma.cshl.org/jian/`