Negative Information for Motif Discovery

K.T. Takusagawa and D.K. Gifford

Pacific Symposium on Biocomputing 9:360-371(2004)

NEGATIVE INFORMATION FOR MOTIF DISCOVERY

K. T. TAKUSAGAWA
kenta@mit.edu

D. K. GIFFORD
gifford@mit.edu

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

We discuss a method of combining genome-wide transcription factor binding data, gene expression data, and genome sequence data for the purpose of motif discovery in *S. cerevisiae*. Within the word-counting algorithmic approach to motif discovery, we present a method of incorporating information from negative intergenic regions where a transcription factor is thought *not* to bind, and a statistical significance measure which account for intergenic regions of different lengths. Our results demonstrate that our method performs slightly better than other motif discovery algorithms. Finally, we present significant potential new motifs discovered by the algorithm.

1 Introduction

In the field of computational biology, motif discovery is one tool for unraveling the transcriptional regulatory network of an organism. The underlying model assumes that a transcription factor binds to a specific short sequence ("a motif") in an intergenic region near a gene the factor regulates. With the recent availabilty of many genome-wide data sets, we can predict certain motifs by computational methods rather than laborious experimentation. Such computational techniques rely on fusing genome sequence data with other data sets. In this paper, we discover motifs by fusing sequence data with transcription factor binding data and gene expression data.

Chromatin immunoprecipitation (ChIP) microarray experiments can determine where in the genome particular transcription factor binds to a resolution of single intergenic region (usually 500-2000 bp)[8]. The GRAM algorithm[2] combines such genome-wide location information with gene-expression experiments. The algorithm discovers additional intergenic regions that are likely bound by the transcription factor but did not cause a strong signal in the ChIP experiment.

For motif discovery, intergenic regions are partitioned into two categories: those to which the transcription factor is thought to bind (according to raw ChIP experiments or after incorporating additional information via an algorithm like GRAM) and those to which it does not bind. We will refer to the bound sequences as the "positive intergenic sequences" and those not bound as the "negative intergenic sequences".

If an algorithm were only to use the positive sequences for motif discovery, then it would likely discover many false motifs. Such false motifs are caused by sequences which appear frequently in all the intergenic sequences of a genome. In *S. cerevisiae*, two prominent simple examples of such sequences are poly-A (long strings of consecutive adenine nucleotides) and poly-CA (long strings of alternating cytosine and adenine nucleotides)[7].

Fortunately, fusing binding data with the complete sequencing of the *S. cerevisiae* genome provides us with a conceptually simple method of discovering a transcription factor's motif: find a sequence which is present in the positive sequences and not present in the negative sequences. However, because of experimental noise and variability of binding by a transcription factor, we expect to find occasional examples of the correct motif in the negative sequences, so we instead seek a motif that is significantly over-represented in the positive intergenic sequences when compared with the negative intergenic sequences.

1.1 Related work

There have been many past efforts to use negative intergenic sequences to derive a statistical test.

The very popular "Random Sequence Null Hypothesis" (so named in Barash, *et al.*[3]) uses the negative sequences to discover the parameters of an n-th order background Markov model ($n = 0$ and $n = 3$ are popular). This approach greatly dilutes the information content of the negative intergenic sequences, and especially loses information about false motifs whose length is greater than the order of the Markov model.

In contrast, the approach pursued in this paper will be similar to Vilo, *et al.*[11] and Barash, *et al.*[3]. Vilo, *et al.* cluster genes by their expression profiles and seek to discover motifs within each cluster. Their test for significance compares the total occurrences of a potential motif in all intergenic sequences to the within-cluster count. Their significance test compares a statistic against a binomial distribution. Barash, *et al.* describe an alternative to the "Random Sequence Null Hypothesis", namely a "Random Selection Null Hypothesis". They perform a similar calculation to Vilo, *et al.*, but compare against a hyper-geometric distribution. (The difference appears to be the assumption of whether motif-containing sequences are selected "with replacement" or "without replacement" from all the sequences.)

A somewhat different approach is described by Sinha[9], who shows how to view motif discovery as a feature selection problem for classification. Sinha's algorithm requires the input of positive and negative intergenic sequences.

Sinha generates the negative examples (intergenic sequences) artificially using a Markov model, but the framework presented the paper could easily use actual negative intergenic sequences from ChIP experiments.

This paper makes the following two contributions to field. First, we describe modification to statistical methods of Vilo, *et al.* and Barash, *et al.* which allow for intergenic sequences with different lengths. Second, we also apply our motif discovery method and statistical test transcription factor binding data from ChIP microarray experiments. The papers cited above were published before ChIP data were available, therefore the authors used clustered gene-expression data for groups of genes thought to be regulated by a common transcription factor.

Recently, other researchers have taken techniques similar those described in this paper and fused them with other data sets. Kellis, *et al.*[6] incorporate conservation information from different yeast species. Gordon, *et al.*[5] incorporate structural data about the transcription factor and its likely binding domain.

2 Methods

We perform motif discovery in the framework of *word-counting*. This framework exhaustively enumerates a class of potential motifs (or *words*) and scores each word for its likelihood of being a true motif. We searched for potential motifs of width 7 with up to 2 wildcard elements among the 7 positions. The wildcard elements permitted were the double-degenerate nucleotides (IUPAC codes M, R, W, S, Y, K) and the quadruple-degenerate "gap" nucleotide (IUPAC code N).

For each potential motif m, we determine which positive sequences and which negative sequences m occurs. We then determine if m occurs in the positive sequences more often then would be expected by chance. We must therefore first define a *null hypothesis* of what in fact is expected by chance. Biologically, the null hypothesis corresponds to the situation that m is not the motif for the transcription factor. To be able to statistically reject the null hypothesis, we must quantify what we would expect to see if the null hypothesis were true. We will present two different null hypotheses, the latter which will incorporate sequence lengths as additional information to the statistical measure.

Computational constraints determined the limits of width 7 and 2 wildcards. At those limits, a search for a transcription factor's motif (within approximately 3 Mbase of *S. cerevisiae* sequence) took approximately 20 minutes on a 1.6 GHz Athlon system. The running time scales exponentially with re-

spect to the width and number of allowed wildcards.

As an aside, we note that this exponential increase could be addressed in future investigations in two ways. For slightly wider motifs or more wildcards, more computing power can be applied: the algorithm parallelizes trivially by having different processors examine separate regions of the search space. Beyond that, if one wanted to discover long motifs, one can use the short motifs discovered by exhaustive search as starting points to an expectation-maximization type algorithm, as done in by Barash, *et al.*[3] and Gordon, *et al.*[5].

2.1 Sequences chosen with uniform probability

The two null hypotheses are instances of the "Random Selection Null Hypothesis" of Barash *et al.*[3], which states that when the null hypothesis is true (i.e., the motif is incorrect), the positive sequences are "randomly selected" from among all the intergenic sequences, without any correlation or bias toward sequences containing the incorrect motif. (One can visualize a transcription factor as the "hand" which randomly selects from an urn of intergenic sequences.) For their model, "randomly selected" means "all sequences are equally likely to be chosen without replacement". For this definition of "randomly selected", they give a formula for the probability that m occurs in k sequences by chance alone.

$$P_{hyper}(k \mid n, K, N) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{1}$$

where n is the number of positive sequences, N is the total number of sequences (positive and negative), and K is the number of sequences in which the word m occurs. The above formula is the hyper-geometric probability distribution.

Using this formula we can calculate a p-value that the null hypothesis is true. The p-value sums the tail of the probability distribution for $k' \geq k$.

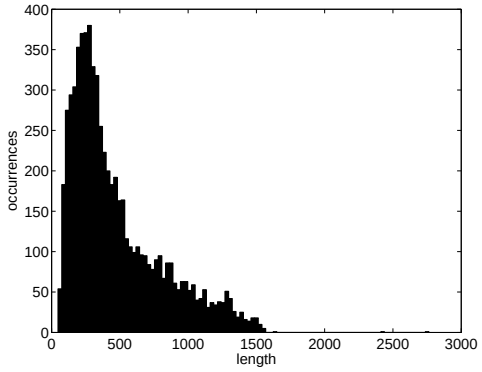
$$\text{p-value}(k) = \sum_{k'=k}^{n} P_{hyper}(k' \mid n, K, N) \tag{2}$$

2.2 Sequences chosen by length

Instead of "all sequences equally likely" as the behavior under the null hypothesis, we propose the null hypothesis that:

> **Sequences will be selected (without replacement) with probability proportional to the sequence's length.**

Figure 1: Distribution of integenic sequence lengths in *S. cerevisiae*.



The motivation for this alternative stems from the fact that sequences from the ChIP experiments are of different lengths (Figure 1). The modification is plausible: given no other knowledge about the transcription factor, a longer sequence is more likely to contain the transcription factor's true motif.

Let AL be the bag (multi-set) of all sequence lengths, and KL be the sub-bag of the lengths of the sequences in which the word m occurs. (Thus $|AL| = N$ and $|KL| = K$.) We use bags to allow for distinct sequences which happen to have the same length.

Having defined the null hypothesis, we can define the probability of it being true as the probability that k or more sequences in which word occurs are selected. Because computing this probability exactly is computationally prohibitive, we instead compute an approximation. Instead of selecting sequences without replacement, we select sequences *with* replacement. The probability of selecting exactly k sequences is binomial:

$$P_{binom}(k \mid n, KL, AL) = \binom{n}{k} r^k (1-r)^{n-k}. \tag{3}$$

where r is the proportion of total sequences (weighted by lengths) containing the word.

$$r = \frac{\sum KL}{\sum AL}$$

To calculate the p-value that the null hypothesis is true, we reuse equation 2, substituting P_{binom} for P_{hyper}.

Table 1: Consensus sequences

TF	Consensus	TF	Consensus
ABF1	TCRNNNNNNACG	CBF1	RTCACRTG
GAL4	CGGNNNNNNNNNNCCG	GCN4	TGACTCA
GCR1	CTTTCC	HAP2	CCAATNA
HAP3	CCAATNA	HAP4	CCAATNA
HSF1	GAANNTTTCNNGAA	INO2	ATGTGAAA
MATa1	TGATGTANNT	MCM1	CCNNNWWRGG
MIG1	WWWWSYGGGG	PHO4	CACGTG
RAP1	RMACCCANNCAYY	REB1	CGGGTRR
STE12	TGAAACA	SWI4	CACGAAA
SWI6	CACGAAA	YAP1	TTACTAA

3 Results and Discussion

The results and discussion are organized into the following sections. §3.1 validates the algorithm by attempting to replicate known motifs. §3.2 presents potential new motifs discovered by the algorithm. Finally, §3.3 discusses ideas for future work.

3.1 Validation

This section measures and compares the algorithm's motif discovery performance. For an absolute measure, the algorithm was run on binding data for transcription factors whose motifs were previously discovered and confirmed biologically. For a comparative measure, the same data were analyzed with the motif discovery programs MEME [1] and MDscan [7]. The algorithm was also run on differently processed binding data for each transcription factor to determine the effect of the type binding data on motif discovery.

Program parameters

MDscan was run through the web interface with the following parameters:

- Motif width: `7`
- Number of top sequences to look for candidate motifs: `10`
- Number of candidate motifs for scanning the rest sequences: `20`
- Report the top final `10` motifs found
- Precomputed genome background model: `S. cerevisiae intergenic`

MEME was run with the command-line parameters `-dna -w 7 -nmotifs 10 -revcomp -bfile $MEME/tests/yeast.nc.6.freq`. The parameters direct MEME attempt to discover 10 motifs of width 7 on either strand using the pre-computed order-6 Markov background model of the yeast non-coding regions.

Binding data

Three different sets of positive sequences were used. That is, three different methods were used to determine which sequences are bound by a transcription factor. The first two are a simple p-value threshold on the ChIP experiment[8] (*not* related to the p-values calculated the statistical tests of Chapter 2). The last uses the GRAM gene modules described in Bar-Joseph, *et al.*[2] which fuse both binding data and expression level data.

1. Bound intergenic regions, cutoff p-value 0.001
2. Bound intergenic regions, cutoff p-value 0.0001
3. GRAM Gene modules under YPD

To score the performance of both this paper's algorithm, and MEME and MD-Scan, the discovered motifs were compared against the consensus sequences for transcription factors (Table 1) which were gathered from the TRANSFAC database.

We score the closeness of a discovered motif with the consensus using a Euclidean distance metric described in the thesis version of this paper[10]. The threshold of correctness was chosen "by eye" to be a value for which discovered motifs below the threshold seemed close to consensus motifs. The threshold was loose enough that a motif is scored "correct" even when the discovered motif spans only half of a wide gapped motif (for example ABF1 or GAL4).

We report the number of times the most statistically significant discovered motif was correct, and the number of times a correct motif was found somewhere in the top 10 significant motifs. This paper's algorithm only reported motifs with significance greater than 10^{-4}, so sometimes no motifs were found. Table 2 gives the number of correct motifs found by the algorithm and other motif-discovery algorithms on different data sets. We can make the following observations:

- The best performance was this paper's algorithm using binding data with threshold p-value 0.001.

Table 2: Verified consensus motifs

Algorithm	Data set	Choose from	Number correct (out of 20)
This paper	p=0.001	Top 10	14
MDscan	p=0.001	Top 10	12
MEME	p=0.001	Top 10	10
This paper	p=0.001	Top 1	10
MDscan	p=0.001	Top 1	9
MEME	p=0.001	Top 1	0
This paper	GRAM	Top 10	12
This paper	GRAM	Top 1	9
This paper	p=0.0001	Top 10	12
MEME	p=0.0001	Top 10	12
This paper	p=0.0001	Top 1	9

- Choosing a more rigorous threshold for the binding data, namely 0.0001, resulted in slightly poorer performance, most likely because of insufficient positive intergenic sequences for a significant result.
- Incorporating gene expression information with the GRAM modules algorithm caused the algorithm to perform slightly poorer than using the raw binding data. However, the modules result did find 2 correct motifs that the raw binding data did not (at the cost of failing to 4 others).
- The algorithm finds slightly more correct motifs than MEME or MDscan.

3.2 New motifs

Tables 3 and 4 give the top-scoring motifs for some transcription factors not listed in Table 1. These are candidates for further investigation. The positive sequences used for the table were the bound sequences at p-value 0.001. From discussion with a colleague, we note that the motifs for CIN5, GAT3, GLN3, IME4, YAP5, and YAP6 are probably not correct, while those for BAS1, FKH1, FKH2, INO4, and SUM1 are consistent with what is known about the transcription factors[4].

Results on shuffled data

To judge the background level of motifs, the algorithm was also run on random sets of intergenic sequences. Ideally, these runs should produce no significant

Table 3: Top scoring motifs discovered for transcription factors not on Table 1 with binomial significance greater than 10^{-10}. The significance values are $\log_{10}$ of the p-value. The gap wildcard is denoted by a dot.

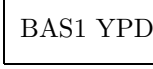
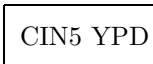
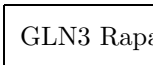
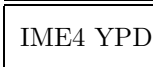
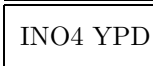
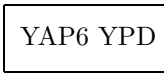
TF + Condition	Motif	Binomial	Hypergeometric
BAS1 YPD		-10.99	-14.71
CIN5 YPD		-10.86	-19.67
FHL1 Rapamycin		-27.28	-39.88
FHL1 YPD		-35.12	-50.61
FKH1 YPD		-10.85	-14.72
FKH2 YPD		-12.16	-18.49
GAT3 YPD		-15.90	-21.14
GLN3 Rapamycin		-11.46	-16.65
IME4 YPD		-12.16	-15.22
INO4 YPD		-12.14	-14.36
MBP1 YPD		-20.14	-25.40
MET4 Rapamycin		-10.25	-13.13
MET4 YPD		-10.78	-13.08

Table 4: Top scoring motifs (continued from Table 3)

TF + Condition	Motif	Binomial	Hypergeometric
NRG1 YPD		-11.65	-19.00
PHD1 YPD		-10.86	-20.01
RGM1 YPD		-12.91	-15.94
STB1 YPD		-10.91	-12.36
SUM1 YPD		-11.38	-17.18
YAP5 YPD		-11.94	-16.98
YAP6 YPD		-11.44	-18.78

motifs. Twenty-five random trials were run for each of 20, 40, 80, 120, and 160 randomly chosen *S. cerevisiae* intergenic sequences (for a total of 125 trials). Five of the 125 experiments discovered a total of 11 motifs with binomial p-values less than 10^{-4}, with most significant motif having significance $10^{-4.7}$. These falsely significant motifs were more likely to be found when there were fewer positive sequences, as 8 of the 11 motifs were found in data sets with 20 positive sequences. In the course of the 125 trials, over 70 million hypotheses (i.e., candidate motifs) are tested, so it is reasonable to see a few false positives with significance has higher than 10^{-4}.

3.3 Future work

The statistical test developed in Chapter 2 can make use of more information for a better measure of significance. In §2.2 we defined the null hypothesis behavior "random selection" to be as selection with probability proportional to length. A straightforward modification would be to instead use the number of different subsequences of a sequences as its probability (appropriately normalized). As an extreme example, consider a very long sequence consisting of a repeat of a single nucleotide. While long, such a sequences offers few possibilities of where a transcription factor might bind. Such a long repetitive

sequence ought to be selected with low probability.

Continuing in this manner, other biological prior knowledge can be incorporated into the prior probability that a sequence is selected. Such knowledge might involve the location of the sequence on the chromosome, knowledge about the gene which the sequence precedes, or other genetic markers.

Biologically, we must question the assumption of independence (modulo choosing without replacement) between the $n = |P|$ random selections from A. For example, it would be reasonable to hypothesize that if two sequences are very similar, they would likely both be selected, or neither.

Not only can we incorporate biologically relevant information into the prior probability of the binding, but we can also try to incorporate more information about the binding event itself. Currently, the algorithm only makes use of the binary presence ("yes" or "no") of words in sequences. It could, for example, incorporate the following features:

- Number of occurrences of the word in the sequence
- Position of the occurrence(s) with respect to the start of transcription or other genetic markers in the sequence
- Strand of the occurrence of the word
- p-value of the binding event.

Beyond yeast, of course, are the many organisms whose genomes have been recently sequenced, including human. It will be only a matter of time before ChIP and other genome-scale location experiments are performed on those organisms. We expect that to do worthwhile motif discovery on larger and more complicated genomes, careful attention will have to be paid to the statistical issues and improvements mentioned above.

Acknowledgements

Special thanks to Richard A. Young, D. Benjamin Gordon, and Ziv Bar-Joseph for help with the data sources used in this project. K.T.T. was supported by a NDSEG/ASEE Graduate Fellowship.

References

1. T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proc. 2nd International Conference on Intelligent Systems for Molecular Biology*, 1994.

2. Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. Computational discovery of gene modules and regulatory networks. *(Submitted for publication)*, 2003.
3. Y. Barash, G. Bejerano, and N. Friedman. A simple hyper-geometric approach for discovering putative transcription factor binding sites. In *Algorithms in Bioinformatics: Proc. First International Workshop*. 2001.
4. D. B. Gordon, 2003. personal communication.
5. D. B. Gordon, L. Nekludova, N. J. Rinaldi, C. M. Thompson, D. K. Gifford, T. Jaakkola, R. A. Young, and E. Fraenkel. A knowledge-based analysis of high-throughput data reveals the mechanisms of transcription factor specificity. *(Submitted for publication)*, 2003.
6. M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423:241–254, 2003.
7. X. S. Liu, D. L. Brutlag, and J. S. Liu. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology*, 20:835–839, August 2002.
8. B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290:2306–2309, 2000.
9. S. Sinha. Discriminative motifs. In *Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, 2002.
10. K. T. Takusagawa. Negative information for motif discovery. Master's project, Massachusetts Institute of Technology, July 2003.
11. J. Vilo, A. Brazma, I. Jonassen, A. Robinson, and E. Ukkonen. Mining for putative regulatory elements in the yeast genome using gene expression data. In *Proc. International Conference on Intelligent Systems for Molecular Biology*, 2000.