Data-Driven Extensions to Gibbs Sampling for Motif Discovery

Matthew D. Edwards

April 21, 2008

Identification of the regulatory regions in the genome is an important task in computational biology. Short sequences of DNA serve as binding targets for a set of proteins, transcription factors, that regulate transcription in the cell. Here we propose using the stable positional preferences of transcription factors relative to transcription start sites ("positional footprinting" or "spatial conservation") to improve DNA motif discovery using ChIP-chip data. Initial results do not show significant improvements using experimental TSS data over an uninformative uniform prior. We propose two extensions to this model that could improve performance. These results highlight the importance of finer-grained sequence-based priors as well as discriminative approaches.

Contents

1 Introduction

Uncovering the pattern of DNA-protein interactions within the cell is an important task, undertaken with the goal of understanding complex networks of regulation and cellular behavior. Identifying protein binding sites is a challenging problem because of the many choices of binding positions within the genome, ambiguous models of motif preference, and the inherent randomness of molecular interaction. Algorithms to solve this problem fall into two main categories: enumerative and statistical methods. Enumerative methods list and score many possible motifs, though the search space must necessarily be pruned and therefore the space of possible motifs is constrained (for instance, see [35] or [36]). Statistical methods use probabilistic models to assign likelihoods for candidate motif patterns and then use various sampling or optimizing techniques to find the most probable patterns and locations (for instance, see [3], [23], or [25]). Over one hundred algorithms and extensions have been proposed; for a more complete overview see the citations in review articles including [18], [21], and [43]. This paper begins by reviewing the motif discovery problem and presenting the PRIORITY framework. Its contributions are three ways of integrating positional information into PRIORITY: the first and simplest shows no global improvement, motivating the introduction of the latter two. The latter two models have not yet been experimentally evaluated.

1.1 Molecular Biology

The central dogma of molecular biology is the flow of information in the cell from DNA to RNA to protein. DNA in the nucleus of a cell serves as a blueprint for RNA, which moves outside the nucleus to be translated into protein. Here we are concerned with transcription factors (TFs), proteins that bind DNA and regulate the process of copying some segment of DNA into RNA (transcription). The exact location where a particular TF binds to DNA is aptly called the transcription factor binding site (TFBS) and is determined by the protein's interactions with the DNA at the TFBS (as well as other factors beyond the sequence). The particular DNA sequence (or group of similar sequences) that a TF binds to is an instance of a motif, a small pattern of nucleotides with biological significance. The position where transcription begins is called the transcription start site (TSS). For some segment of DNA (a gene) to be transcribed into RNA, several TFs must assemble upstream of the gene and recruit other

proteins (including RNA polymerase) to begin transcription. Some TFs act as repressors or recruit other, related, proteins for different purposes and thus can act at varying distances.

Once an RNA copy of the gene has been transcribed, the cell performs some alterations on it and exports it out of the nucleus. Outside the nucleus, the ribosome binds to the RNA and begins translating it into a peptide, using a three-nucleotide code. An important distinction is the difference between the TSS and the translation start site: the ribosome begins translation at the translation start site, AUG. Some number of RNA bases are ignored before this (called the 5′ untranslated region, UTR), making the *transcription* start site location farther upstream than the *translation* start site. In this paper TSS always refers to the transcription start site.

1.2 Motif Discovery

A common formulation of the motif discovery problem is to predict regulatory motifs and sites given a set of DNA sequences that have been experimentally shown to bind to a specific TF ([38], [45]). Data of this type arise from many kinds of experiments, including gene expression arrays ([39], [22]), ChIP-chip experiments ([20], [26]), and DNA-binding arrays ([29]).

1.3 Binding Model

The model of DNA-protein binding is a key aspect for motif discovery algorithms. A common model to describe a binding protein's affinity for specific DNA sequences is a position-specific scoring matrix, or PSSM, that at each position describes the probability of binding to each of the four possible nucleotides ([40]). This model assumes independence between positions in a motif, which has been shown to be a suboptimal ([1], [4], [6], [49]) but still decent ([5]) approximation. For a graphical representation of a DNA binding motif, see figure 1.

The PSSM allows us to compute the probability of a motif binding to some particular DNA sequence.

1.4 PRIORITY framework

PRIORITY is a motif discovery algorithm designed to use relevant information about the probability that TFs bind to certain parts of a sequence ([17],

Figure 1: Transcription factor PDR1 motif reported in [20], visualized using [10]. Each letter's height represents the preference for a given nucleotide at a given position. The scale measures entropy to reflect the information content of a distribution.

[30], [31]). The PRIORITY model takes a set of n sequences, X_1 to X_n, that are believed to bind to a particular TF. The length of sequence X_i is m_i and $X_{i,j}$ identifies the base at position j in sequence i. For simplicity, we assume there is either zero or one binding target in each X (the zero-or-one-occurrence, ZOOPS, model from MEME ([3])). The zero binding site option accounts for false positives in the underlying experiments. To specify the binding sites we define a vector Z with $Z_i = j$ indicating that the binding site in sequence i starts at position j. We define $Z_i = 0$ to mean that sequence i has no target site. PRIORITY includes the non-motif background nucleotides with some model parametrized by ϕ_0 and the motif (of length W) by ϕ. Currently, the motif model is a PSSM specified by a matrix ϕ where $\phi_{a,b}$ is the probability of finding base b at position a. Therefore, the likelihood of observing some set of sequences when we know the hidden, explanatory states is:

$$\Pr(X_i|\phi, Z_i > 0, \phi_0) =$$
$$\Pr(X_{i,1}, \ldots, X_{i,Z_i-1}|\phi_0) \left(\prod_{k=1}^{W} \phi_{k,X_{i,Z_i+k-1}} \right) \Pr(X_{i,Z_i+W}, \ldots, X_{i,m_i}|\phi_0).$$

The joint distribution over all nucleotides in X_i factors into those that are in the background and those that are in the motif, controlled by the PSSM. If the sequence doesn't contain a binding site, every nucleotide comes from the background distribution:

$$\Pr(X_i|\phi, Z_i = 0, \phi_0) = \Pr(X_{i,1}, X_{i,2}, \ldots, X_{i,m_i}|\phi_0).$$

In this inference problem we want to find the most likely hidden states (ϕ and Z) given the observed sequences (X). So our objective function is:

$$\underset{\phi,Z}{\operatorname{argmax}} \Pr(\phi, Z|X, \phi_0) = \underset{\phi,Z}{\operatorname{argmax}} \left\{\Pr(X|\phi, Z, \phi_0) \Pr(\phi_0) \Pr(\phi) \Pr(Z)\right\}.$$

1.5 Optimization via Sampling

We choose to run a Gibbs sampler ([15]) to sample the posterior distribution and hopefully visit maximal values of ϕ and Z. The derivation of a predictive update step is also in [30] and later papers, all based on the collapsed Gibbs sampler ([24]). Gibbs sampling is a Monte Carlo Markov Chain (MCMC) algorithm that makes a Markov chain whose stationary distribution is the true joint distribution. It allows us to generate samples that converge to some distribution by using only conditional distributions (for details, see [8]). It is useful in cases like this where the complete distribution (of Z and ϕ) is too complicated to optimize or sample from, but the conditional distributions are simple. That is, we know the probability of a PSSM from known binding sites, and conversely, we can compute binding site probabilities from the PSSM. We use collapsed Gibbs sampling ([24]) to skip the expensive step of sampling from ϕ and to sample each Z_i from only other Z:

$$\Pr(Z_i|Z_{[-i]}, X, \phi_0) = \frac{\Pr(Z|X, \phi_0)}{\Pr(Z_{[-i]}|X, \phi_0)} = \frac{\Pr(Z) \int_\phi \Pr(X|\phi, Z, \phi_0) \Pr(\phi) d\phi}{\Pr(Z_{[-i]}) \int_\phi \Pr(X|\phi, Z_{[-i]}, \phi_0) \Pr(\phi) d\phi},$$

where $Z_{[-i]}$ means the vector of binding sites Z without Z_i. As in [24], we can successfully integrate out ϕ when its prior is Dirichlet. We also divide out $\Pr(Z_i = 0, X_i|\phi_0)$ which is a constant at each sampling step. We obtain a "predictive update" distribution over each j with a particular sequence X_i, given the remaining binding sites $Z_{[-i]}$:

$$\Pr(Z_i = j|Z_{[-i]}, X, \phi_0) = \frac{\Pr(Z_i = j) \left(\prod_{a=1}^{W} \phi_{a, X_{i,j+a-1}}\right)}{\Pr(Z_i = 0) \Pr(X_i, \ldots, X_{i,j+W-1}|\phi_0)}$$

for $j > 0$ and

$$\Pr(Z_i = 0|X, \phi_0) = 1.$$

Note that ϕ is computed from the nucleotide composition at each binding site $Z_{[-i]}$ as well as smoothing pseudocounts from the Dirichlet prior. The

posterior probability of the joint distribution is then:

$$\Pr(\phi, Z|X, \phi_0) \propto \Pr(X|\phi, Z, \phi_0) \Pr(\phi_0) \Pr(\phi) \Pr(Z).$$

PRIORITY, in contrast to other models, employs a non-uniform prior over the sequences, $\Pr(Z)$. The Gibbs sampler generates many samples from the posterior distribution, and we only store the most likely as our reported motif. However, we could use the entire sampled distribution to find multiple motifs or compute a centroid estimate (as in [32]) for a possibly improved motif. The PRIORITY model isn't necessarily constrained to Gibbs sampling for optimization or even a PSSM binding model; more complex optimization strategies or binding models could be implemented.

2 Single TSS Prior

As shown in other versions of PRIORITY, informative priors on the DNA sequence sets can improve motif discovery. We might expect there to be some organized pattern between the binding site of a TF and the associated TSS, perhaps arising from the biochemical nature of its regulation or the behavior of the various complexes it recruits to begin transcription. Certain TFs have been observed to have specific distance preferences relative to the TSS in yeast ([16], [34], and [47]) and humans ([46]), but the literature does not extend to all TFs. At a larger level, groups of similar positional distributions (which we will discuss later) have been observed in yeast ([13]).

2.1 History

Positional preferences have been observed and used at various stages to aid motif discovery. A program was developed to analyze motif distance distributions ([37]), and it identified significantly nonrandom distributions for several yeast TFs. A web tool to visualize and analyze positional clustering has been developed ([7]), which the authors used to observe several consistent relationships and coined the phrase "positional footprinting." These programs were built to analyze relationships with known motifs and binding sites. Other tools have been built to find binding sites from known motifs ([44] and [11], who apparently introduced the phrase "spatial conservation"). These use various statistical techniques to identify sufficiently nonrandom clusters of matching

patterns and identify functional binding sites.

Closer to what we propose, a consistent positional relationship relative to the TSS alone has been used to propose novel motifs in humans ([14]). The authors enumerated all 8-nucleotide motifs and used a simple variance criterion to measure a TF's clustering. Improbizer (see supplementary methods of [2]) includes an option to bias the distribution of binding sites to a Gaussian pattern, but implicitly requires a uniform alignment to the TSS for each promoter region and performs a greedy optimization instead of sampling. Explicit TSS alignment is leveraged in [41] in a combined enumerative/sampling framework. Similarly, a TSS weighting option is mentioned briefly in the "Gibbs Recursive Sampler" ([42]), but was derived from prokaryotic data and translation start sites.

We seek to explore and exploit this relationship between TFs and start sites within our PRIORITY framework. In one sense, we will add overrepresentation analysis and flexible PSSM models to the earlier work in [14]. In another sense, we will use the intuitions that motivated the analysis in [7], [13], [37], and [47] in reverse to help find motifs. Our work extends that of [2], [41], and [42] in that it incorporates varying locations (and sometimes number and confidence) of start sites from experimental data in a more robust statistical sampling framework. We will use this start site distance preference to build a positional prior, which we call PRIORITY-$\mathcal{S}$.

2.2 Experimental Data

To discover the distance characteristics of TFBS-TSS interaction in aggregate, we start with transcription start sites reported in two experiments. Zhang and Dietrich ([48]) used 5′ SAGE to identify the exact location of the start of mRNA transcripts in 2231 yeast genes. Miura and collaborators ([28]) performed full cDNA sequencing with vector-capping to identify transcription start sites for 3599 yeast genes.

Notably, most genes had two or more reported start sites. As stated in the later paper, 1693 genes had hits in both studies but only 41% shared a reported start site. This fact suggests that transcription initiation is not wholly deterministic or unique and that relying on a single, canonical start site for each gene might be a poor approximation.

2.3 Calculation of Prior

We combine these reported start sites with 4387 high-confidence binding site locations from [27] and generate a histogram of the distances between each binding site and its closest start site in both directions. We see that the overall distribution for all TFs is markedly different than a random distribution (figure 2), concordant with the results in [20] for TF-translation start site distances.

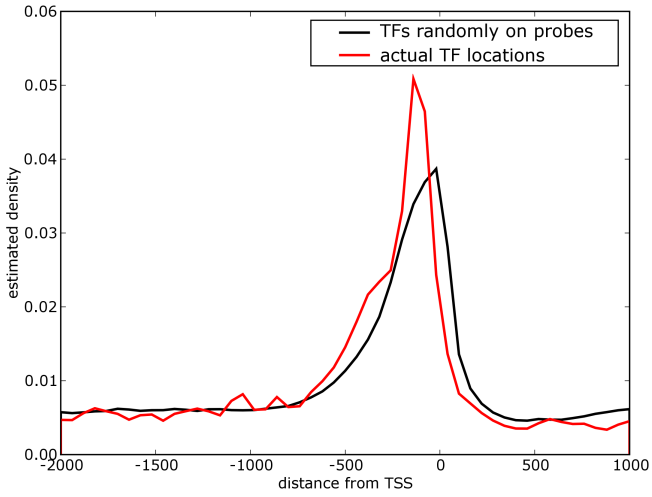

Figure 2: TSS-TFBS distance from observed data and from a randomized distribution

We now wish to transform this distribution into a series of probabilistic scores that correspond at a single-nucleotide level to the probability that a binding site starts at each position. We use a kernel density estimator (Parzen window method, see [33]) to non-parametrically extend this histogram to every position.

This unscaled distribution describes the distance preference for binding sites at particular distances from transcription start sites. However, not all yeast genes have experimentally reported transcription start sites. To generate probabilistic scores for the promoter regions of those genes, we need a distribution to describe the distance preference of TFs from translation start starts (as opposed to transcription start sites). To do this we convolve our TF-TSS distribution with a TSS-translation start site distribution to generate a TF-translation start site distribution. We obtain the TSS-translation start site from the TSS sites

and known translation start sites from SGD ([9] and updates).

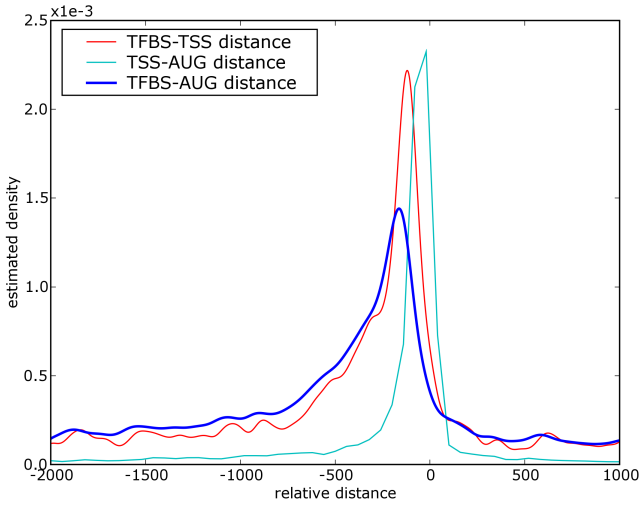

Figure 3: TFBS-translation start site distribution obtained by convolving the empirical TSS-TFBS and TSS-TFBS distributions

We now wish to extend this distribution for a single TF and start site to a probabilistic score we can assign to each nucleotide, reflecting the belief that it begins a binding site regulating any gene. We use the notation $\Pr(i \in \text{TFBS})$ to mean the probability that the W-mer starting at position i on some chromosome is a binding site, and $\Pr(i \text{ regulates } j)$ to mean that it regulates transcription starting at j on the same chromosome. The function $f(d)$ is proportional to the probability that a binding site is distance d from its targeted start site. The distribution in figure 2 is $f(d)$ for experimental start sites and the convolved distribution in figure 3 is $f(d)$ for translation start sites without experimental data. Now we can compute $\Pr(i \in \text{TFBS})$ based on quantities we know:

$$\Pr(i \in \text{TFBS}) = 1 - \Pr(i \notin \text{TFBS}) = 1 - \prod_j (\Pr(i \text{ not regulates } j, j \in \text{TSS}) + \Pr(j \notin \text{TSS}))$$

$$= 1 - \prod_j (\Pr(i \text{ not regulates } j | j \in \text{TSS}) \Pr(j \in \text{TSS}) + \Pr(j \notin \text{TSS}))$$

$$\approx 1 - \prod_j (1 - f(i-j) \Pr(j \in \text{TSS})).$$

Here $\Pr(j \in \text{TSS})$ is our confidence in the experimental data, (arbitrarily) a function of the number of reported transcripts beginning at j divided by the total number of reported transcripts for a gene.

As in [17] and [31], we turn this unscaled probabilistic score for each position being an actual binding site into a valid probability distribution over all positions j in a sequence X_i, that is, $\Pr(Z_i = j)$. We denote our probabilistic score as $S_{a,i}$, reflecting the belief that position i begins a binding site in sequence a. If X_a does not have a binding site, then $Z_a = 0$ and:

$$\Pr(Z_a = 0) \propto \prod_{u=1}^{m_a - W + 1} (1 - S_{a,u}).$$

If X_a has a binding site at a permissible i, then:

$$\Pr(Z_a = j) \propto S_{a,i} \prod_{u=1, u \neq j}^{m_a - W + 1} (1 - S_{a,u}).$$

We see the general characteristics of this new prior in figure 4, an example region from chromosome 7 covering several intergenic regions.

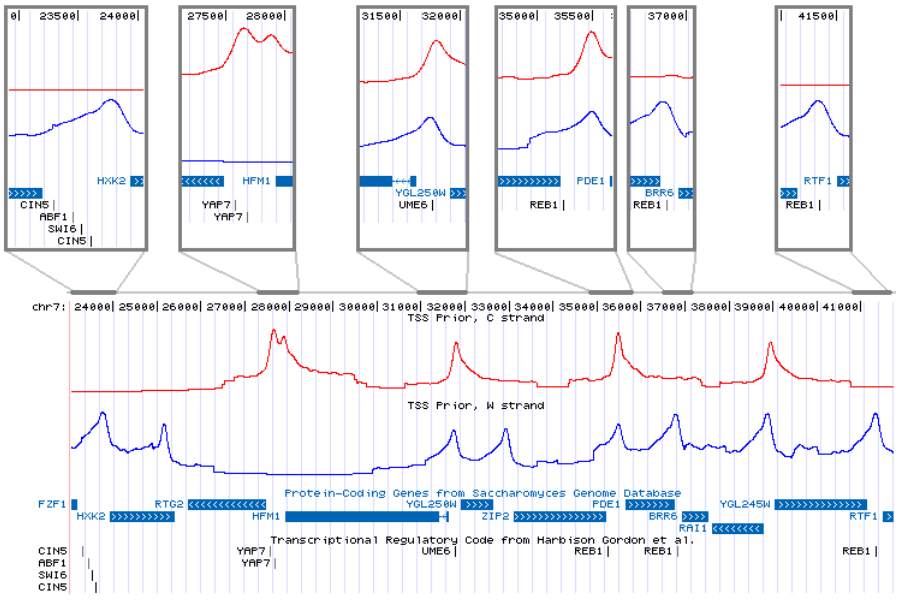

Figure 4: TSS prior on 20kb of chromosome 7

2.4 Evaluation of Motif Discovery Algorithms

Evaluating the performance of an algorithm designed to make conclusions from biological data is a difficult task, since the correct answer is often unknown or ambiguous. Algorithms for motif discovery are sometimes tested on artificially generated data designed to show off some specific advantage, but we feel this approach is not ideal for our model. PRIORITY uses priors designed to model some underlying aspect of the biology (here, positional preferences) that are difficult to artificially generate to the extent that they accurately reflect real data. Instead, we evaluate our performance using the ChIP-chip data of Harbison et al. ([20]), who identified binding regions for 203 yeast TFs in multiple environmental conditions. Of the 242 sequence sets with at least 10 sequences (148 TFs), 156 sets (80 TFs) have known motifs in the literature. We run PRIORITY on these sequence sets and see if the reported motif matches the known value. We compare our reported motif to the literature value, based on a simple squared-error motif similarity metric (see [30] and the supplementary methods section of [20]). More advanced methods of motif similarity exist (for instance, [19]), but this one has been used previously and has the advantage of simplicity for a simple binary decision problem (as opposed to using it for some total error measure).

To evaluate PRIORITY, we run 20 trials of 10,000 iterations each and take the highest-scoring motif from all trials as our reported result. We then compare this to the literature value, and sum the number of correct answers across all sequence sets. Since PRIORITY uses a random algorithm, obtaining the correct motif for a given sequence set is not a deterministic process. When we run 20 trials, we are giving the algorithm 20 chances at a random draw of success. In practice we see a range of correct results between runs, so here we run 5 times (20 trials each) to get an estimate of the reliability and repeatability of the sampler.

2.5 Results

We find no difference in performance between PRIORITY-$\mathcal{S}$ and a uniform, uninformative prior.

Uniform	PRIORITY-$\mathcal{S}$
59.8 ± 0.8	59.4 ± 1.5

However, we do notice that several TFs (in a particular environment) are correctly reported more often using our new prior. We find that these sequence sets are most (differentially) benefited:

Factor	Uniform Runs	Priority-$\mathcal{S}$ Runs
DAL82_SM	1/5	4/5
MAC1_H2OHi	0/5	5/5
MSN2_Acid	1/5	4/5
STB5_YPD	1/5	5/5
ZAP1_YPD	2/5	4/5

Cursory examination shows that these TFs are not particularly unusual: their positional distributions are fairly peaked and they show little downstream regulation. Perhaps the benefits of PRIORITY-$\mathcal{S}$ only apply to certain factors, an idea we explore in the next section.

3 Multi-class TSS Priors

3.1 Motivation

We find that PRIORITY-$\mathcal{S}$ does not improve performance across all TFs. This could be due to the aggregation step used to compute the TSS-TFBS distribution, as many different classes of distance preferences were merged into a single, average distribution. Perhaps there are several distinct classes of distance distributions, which we could separate to provide greater specificity in our positional prior. We will explore a way to find groups of positional classes, with the aim of extending PRIORITY-$\mathcal{S}$.

3.2 Positional Classes and Clustering

Following our earlier approach, we compute TSS-TFBS distributions using high-confidence binding sites from [27]. Now, we seek to divide them into similar groups based solely on positional preferences. To accomplish this, we use hierarchical agglomerative clustering with average linkage. Our similarity metric between distributions is a simple L_2 norm. Other choices for clustering or variants on this method (linkage and distance metrics) are available; these choices were made to simplify intuition for an initial analysis. We apply it to

29 frequently-occurring TFs (combining across environmental conditions), and visualize the clustering in figure 5.

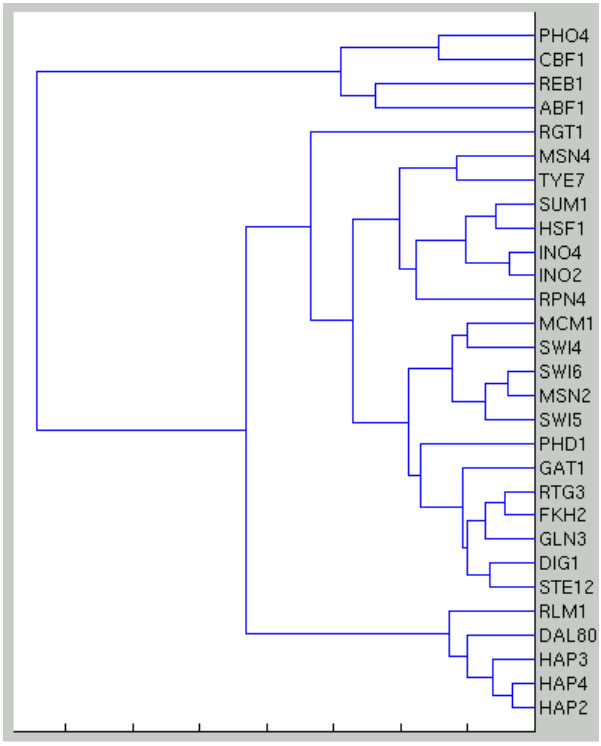

Figure 5: Estimated TF families, computed by hierarchical clustering

Similar grouping of positional distributions was performed in [13], but their analysis was restricted to TFs in unambiguous promoter regions and their clustering method was not stated (perhaps manual). We find that we match the majority of their results; the discrepancy probably arises from binding site prediction (MotEvo vs. the methods of [27]) and dataset size (here we use all promoters and multiple TSSs). To visually assess the properties of the clustering, we analyze the distributions of four groups of factors (RGT1 is omitted) in figure 6.

We see that the clustering method did split factors based on observable positional preference differences. The distributions qualitatively seem to resemble normal distributions, which motivates a type of low-dimensional compression to visualize the distribution families. We estimate the mean and variance of a Gaussian from the interpolated density and visualize these pairs, separat-

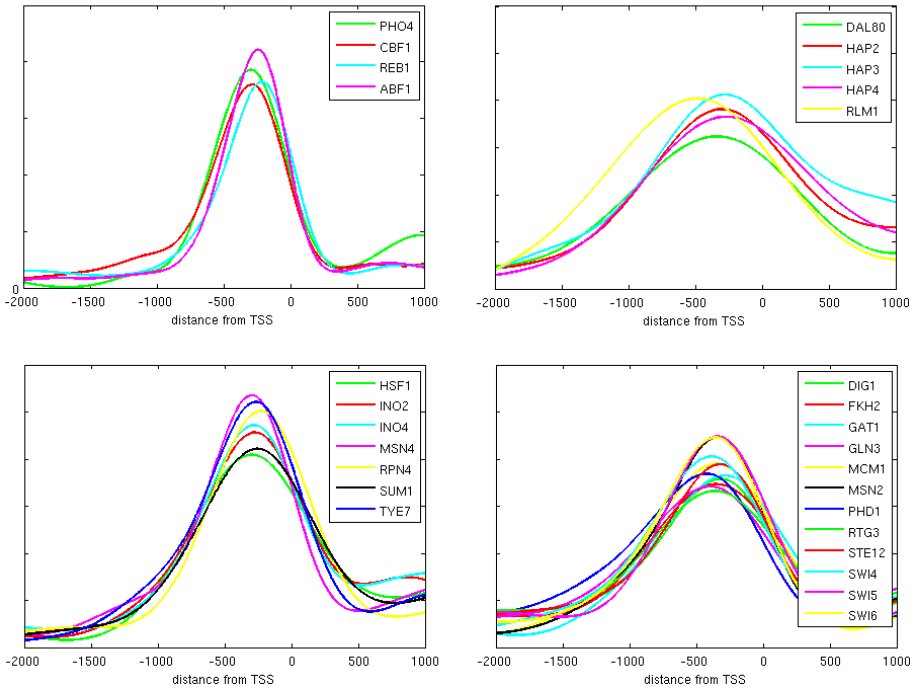

Figure 6: Positional distributions, grouped by cluster

ing previously-obtained clusters by color (figure 7).

Visually, the parameters of the estimated normal distributions retain the differences between clusters. It will be interesting to consult the literature and determine if these artificial classes correspond to known differences in TF type (function, behavior, targets) in the future.

3.3 Integration into PRIORITY

These separate families of positional preferences can be integrated into PRIORITY, as in [30]. We can use the clustering to compute n positional distributions, build n genome-wide priors as described earlier, and extend our sampler to also sample which positional class a factor belongs to.

However, this model has several unappealing aspects. We have not established a biological justification for multiple positional classes, and the inclusion of more and more site data (more separate positional distributions) could suggest that we are "pre-loading" our algorithm with less and less obscured (as n

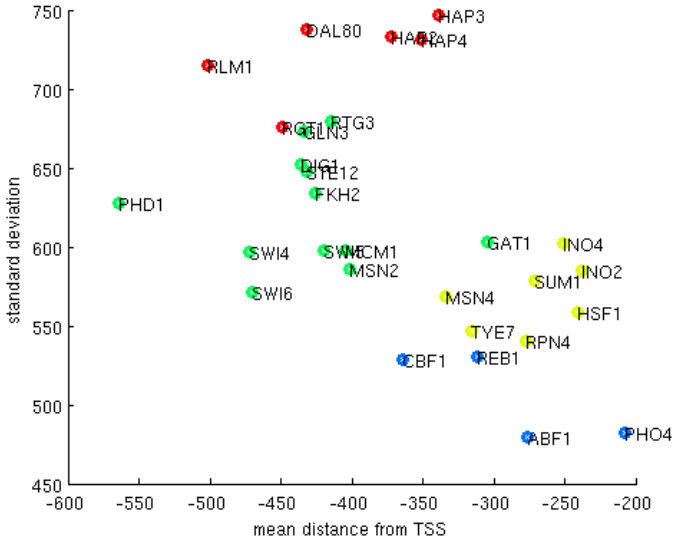

Figure 7: Positional clusters and estimated normal distribution parameters

increases) representations of correct answer. Finding a biological justification for these separate classes, hypothetically some structural difference or varying criticality of function, could alleviate the first problem. The second problem is harder to overcome, since we do need a source of binding site locations to compute our position distributions. This motivates our last model.

4 Dynamic TSS Prior via Hyperpriors

4.1 Motivation

The multi-class model used several positional distributions, largely resembling Gaussians with differing means and variances. In that formulation, each class was generated based on known data, so the number of classes should be kept small so that the right answers are not "pre-loaded" into the prior. Here we propose a model that eliminates a dependence on known binding sites and also allows for an arbitrary number of positional classes. It follows from our observations that the positional distributions of separate classes resemble normal distributions. Here we use a "functional prior" to impose a particular distribution on the TSS-TFBS distances and use Gibbs sampling to estimate the

parameters of that distribution.

4.2 Sampling Details

Our current algorithm computes Z and ϕ but uses collapsed Gibbs sampling to integrate out ϕ and thus only samples Z. We can augment the sampler to learn μ and σ, the parameters of the TSS-TFBS distance distribution. The sampler will now iterate between drawing:

1. $\Pr(Z_i|Z, X, \phi_0, \mu, \sigma)$ for all i
2. $\Pr(\mu|Z, X, \phi_0, \sigma)$
3. $\Pr(\sigma|Z, X, \phi_0, \mu)$.

In effect, the $\Pr(Z)$ term is now dynamic (parametrized by μ and σ, which vary) instead of static as in other versions of PRIORITY. A $\Pr(Z_i = j|\mu, \sigma)$ term arises for the first computation, which is simple to compute for a single TSS (drawing from a normal with known parameters). We sample an unknown mean from the samples of a normal with known variance, simplified by putting a normal hyperprior on the mean (sampling from a normal). Similarly, we compute an unknown variance from samples and a known mean, which is again simple if the variance is distributed according to a inverted gamma distribution (sampling from an inverted gamma).

4.3 Extensions

Several extensions to this preliminary model are available. Different distributions on the TSS-TFBS distance could be employed, as well as different families of hyperpriors on these distributions. Sampling from these altered posterior distributions might be more complicated if we do not use conjugate priors, but possible, especially if we allow approximations (discretization or approximate sampling). We might be able to collapse out the functional prior if the hyperprior is of an acceptable form, but this has not been fully pursued.

Tuning the distribution families and their hyperparameters requires more empirical study and biological justification. Incorporating multiple ambiguous start sites (or handling genes with no reported start sites) into this model introduces further complexity. An intriguing possibility is using this framework to assign probabilities to reported start sites and compare different experimental methodologies, but this remains unexplored.

5 Discussion

In this paper we presented and evaluated PRIORITY-$\mathcal{S}$, an informative positional prior designed to bias TFBS distributions towards consistent distances from transcription start sites. However, it did not show performance gains relative to a uniform prior. This motivated the introduction of two extensions to this algorithm, which have yet to be experimentally evaluated.

Our single-prior model, PRIORITY-$\mathcal{S}$, did not show significant improvements across all sequence sets. This "overall improvement" criterion is the toughest to satisfy; perhaps this prior improves convergence speed, success likelihood, or some other sampling characteristic. It is also possible that the 20 trials and 10,000 iterations we ran masked any improvement over the uniform prior. Improvements of this kind are important, but not as striking and valuable as improving the overall number of motifs found. In the future we will examine the results for subtle improvements to sampling performance. We did, however, identify a small set of factors that only this prior could find, which shows a hint of potential future success. More specific distance distributions, either through our proposed multi-class model or functional prior model, could expand this set of benefited sequence sets.

Several issues remain in our implementation of PRIORITY-$\mathcal{S}$ and the proposed extensions. First, our general practice of learning positional distributions and applying them to find motifs might be theoretically unsound. The positional distribution obtained from real binding site locations is a compressed representation of the desired output, which our algorithm then recovers. To remedy this, we could create a separate distribution for each sequence set by removing the binding sites of the factor in question. This would remove at most 4% of the 4387 sites we used to build the distribution, which is why we omitted this step earlier. However, this "cheating percentage" would grow linearly in the number of classes used for our multi-class prior, and this step would become necessary.

Secondly, the way we incorporate experimental TSS locations into the model could be improved. The data report multiple sites per gene, and the two experiments were in agreement only 41% of the time. Either start sites are highly variable and in many locations per gene, or the experimental data are noisy and unreliable (or both). Both choices necessitate the adoption of ways to incorporate many start sites per gene and varying confidence levels in a principled manner. PRIORITY-$\mathcal{S}$ did both of these, albeit with an ad-hoc TSS confidence

function. This could be extended to the multi-class model, but a similar extension for the functional prior framework requires more work.

Finally, our linear model of TSS-TFBS interaction is an oversimplification of actual DNA behavior. DNA winds around nucleosomes in the cell, which affect the accessibility of certain positions to TFs. Nucleosome occupancy affects the winding of DNA and therefore the notion of distance from a TSS. According to [13], approximately half of the binding sites in yeast are within 145 base-pairs of the TSS (the amount of DNA wound around one nucleosome). Integrating nucleosome occupancy information (as in [30]) with TSS distances could bring the TSS-TFBS interaction model closer to biological reality, particularly for the other half of binding sites.

5.1 Future Work

Several directions of work remain to be done. First, experimental evaluation of the multi-class distance and functional priors is required. Secondly, variation of the several design choices should be performed (number of classes, distribution families, hyperprior forms and parameters, TSS confidence functions). Finally, we can investigate ways to combine this prior with other sources of data, including nucleosome occupancy or discriminative approaches.

For many other informative priors in the PRIORITY framework, we argue that the utility of the prior would increase or at least be maintained for motif-finding in more complex organisms. However, a simple TSS distance prior, constant across all factors, will probably not be as useful as transcriptional complexity increases. This is because the interaction distance dramatically increases in higher organisms ([12]). Multi-class or functional priors might maintain utility, since patterns still do exist in motif distributions of higher organisms ([14], [46]).

5.2 Conclusion

We have proposed and evaluated a single TSS distance prior that shows improvements for several TFs, but not a global improvement over a uniform prior. We have also proposed two extensions to this model that could benefit performance. Until these models are implemented and evaluated, it remains an open question whether or not positional clustering relative to TSS can improve motif discovery on a large scale.

Acknowledgments

I would like to thank my advisor, Alex Hartemink, for his guidance, encouragement, and generosity during this work and his positive influence on my entire undergraduate career. I thank Raluca Gordân and Lee Narlikar for helpful ideas and discussions that started, sustained, and improved much of my work on this problem. Finally, I'd like to thank my friends and family for their support in many adventures in the past and many more to come.

References

[1] Agarwal, P. and Bafna, V. (1998) Detecting non-adjacent correlations within signals in DNA, *Proceedings RECOMB 1998*, 2-8.

[2] Ao, W., Gaudet, J., Kent, W., Muttumu, S., and Mango, S. (2004) Environmentally Induced Foregut Remodeling by PHA-4/FoxA and DAF-12/NHR, *Science*, 305(5691):1743-1746.

[3] Bailey, T. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proceedings ISMB*, 2:28-36.

[4] Barash, Y., Kaplan, T., Friedman, N., and Elidan, G. (2003) Modeling dependencies in protein DNA binding sites, *Proceedings RECOMB 2003*, 2837.

[5] Benos, P., Bulyk, M. and Stormo, G. (2002) Additivity in proteinDNA interactions: how good an approximation is it? *Nucleic Acids Research*, 30(81):4442 4451.

[6] Bulyk, M., Johnson, P., and Church, G. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors, *Nucleic Acids Research*, 30(79):12551261.

[7] Bellora, N., Farre, D., and Alba, M. (2007) PEAKS: Identification of regulatory motifs by their position in DNA sequences, *Bioinformatics*, 23(2):243-244.

[8] Casella, G. and George, E. (1992) Explaining the Gibbs sampler, *The American Statistician*, 46(3):167-174.

[9] Cherry, P. et al. (1997) Genetic and physical maps of Saccharomyces cerevisiae, *Nature*, 387(6632S):67-73.

[10] Crooks G., Hon G., Chandonia J., and Brenner S. (2004) WebLogo: A sequence logo generator, *Genome Research*, 14:1188-1190.

[11] Defrance, M. and Touzet, H. (2006) Predicting transcription factor binding sites using local over-representation and comparative genomics, *BMC Bioinformatics*, 7:386.

[12] Dobi, K. and Winston, F. (2007) Analysis of transcriptional activation at a distance in Saccharomyces cerevisiae, *Molecular and Cellular Biology*, 27(15):5575-5586.

[13] Erb, I. and van Nimwegen, E. (2006) Statistical features of the yeast's transcriptional regulatory code, *IEEE Proceedings ICCSB 2006*.

[14] FitzGerald, P., Shlyakhtenko, A., Mir, A., and Vinson, C. (2004) Clustering of DNA sequences in human promoters, *Genome Research*, 14:1562-1574.

[15] Gelfand, A. and Smith, A. (1990) Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, 85(410):398-409.

[16] Goffeau, A. et al. (1996) Life with 6000 genes, *Science*, 274:546-567.

[17] Gordân, R. and Hartemink, A. (2008) Using DNA duplex stability information for transcription factor binding site discovery, *Proceedings PSB 2008*, to appear.

[18] GuhaThakurta, D. (2006) Computational identification of transcriptional regulatory elements in DNA sequence, *Nucleic Acids Research*, 34(12):3585-3598.

[19] Gupta, S., Stamatoyannopoulos, J., Bailey, T., and Noble, W. (2007) Quantifying similarity between motifs, *Genome Biology*, 8(2):R24.1-R24.9.

[20] Harbison, C., Gordon, D., Lee, T., Rinaldi, N., MacIsaac, K., Danford, T., Hannett, N., Tagne, J., Reynolds, D., Yoo, J., Jennings, E., Zeitlinger, J., Pokholok, D., Kellis, M., Rolfe, P., Takusagawa, K., Lander, E., Gifford, D., Fraenkel, E., and Young, R. (2004) Transcriptional regulatory code of a eukaryotic genome, *Nature*, 431:99104.

[21] Hu, J., Li, B., and Kihara, D. (2005) Limitations and potentials of current motif discovery algorithms, *Nucleic Acids Research*, 33(15):4899-4913.

[22] Kim, S., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J., Eizinger, A., Wylie, B., and Davidson, G. (2001) A gene expression map for Caenorhabditis elegans, *Science*, 293:20872092.

[23] Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A., Wootton, J. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, *Science*, 262:208214.

[24] Liu, J. (1994) The collapsed Gibbs sampler in Bayesian computations with application to a gene regulation problem, *Journal of the American Statistical Association*, 89(427):958-966.

[25] Liu, J., Neuwald, A., Lawrence, C. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies, *JASA*, 90:11561170.

[26] Liu, X., Noll, D., Lieb, J., and Clarke, N. (2005) DIP-chip: Rapid and accurate determination of DNA binding specificity, *Genome Research*, 15(3):421427.

[27] MacIsaac, K., Wang, T., Gordon, D., Gifford, D., Stormo, G., and Fraenkel, E. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae, *BMC Bioinformatics*, 7(113).

[28] Miura, F., Kawaguchi, N., Sese, J., Toyoda, A., Hattori, M., Morishita, S. and Ito, T. (2006) A large-scale full-length cDNA analysis to explore the budding yeast transcriptome, *Proceedings of the National Academy of Sciences*, 103(47):17846-17851.

[29] Mukherjee, S., Berger, M., Jona, G., Wang, X., Muzzey, D., Snyder, M., Young, R., and Bulyk, M. (2004) Rapid analysis of the DNA binding specificities of transcription factors with DNA microarrays, *Nature Genetics*, 36(12):13311339.

[30] Narlikar, L., Gordân, R., Ohler, U., and Hartemink, A. (2006) Informative priors based on transcription factor structural class improve de novo motif discovery, *Bioinformatics*, 22(14):e384-e392.

[31] Narlikar, L., Gordân, R., and Hartemink, A. (2007) Nucleosome occupancy information improves de novo motif discovery, *Proceedings RECOMB 2007*, 4453:107-121.

[32] Newberg, L., Thompson, W., Conlan, S., Smith, T., McCue, L., and Lawrence, C. (2007) A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction, *Bioinformatics*, 23(14):1718-1727.

[33] Parzen, E. (1962) On estimation of a probability density function and mode, *The Annals of Mathematical Statistics*, 33(3):1065-1076.

[34] Pelechano, V., Garcia-Martinez, J., and Perez-Ortin, J. (2006) A genomic study of the inter-ORF distances in Saccharomyces cerevisiae, *Yeast*, 23(9):689-699.

[35] Pesole, G., Prunella, N., Liuni, S., Attimonelli, M., Saccone, C. (1992) WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences, *Nucleic Acids Research*, 20:28712875.

[36] Pevzner, P. and Sze, S. (2000) Combinatorial approaches to finding subtle signals in DNA sequences, *Proceedings ISMB 2000*, 8:269278.

[37] Quandt, K., Grote, K., and Werner, T. (1996) GenomeInspector: Basic software tools for analysis of spatial correlations between genomic structures within megabase sequences, *Genomics*, 33(2):301-304.

[38] Siggia, E. (2005) Computational methods for transcriptional regulation, *Current Opinion in Genetics and Development*, 15:214-221.

[39] Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher,B. (1998) Comprehensive identification of cell cycleregulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Molecular Biology of the Cell*, 9:32733297.

[40] Staden, R. (1984) Computer methods to locate signals in nucleic acid sequences, *Nucleic Acids Research*, 12:505-519.

[41] Tharakaraman, K., Mariño-Ramirez, L, Sheetlin, S., Landsman, D., and Spouge, J. (2005) Alignments anchored on genomic landmarks can aid in the identification of regulatory elements, *Bioinformatics*, 21(S1):i440-i448.

[42] Thompson, W., Rouchka, E, and Lawrence, C. (2003) Gibbs recursive sampler: finding transcription factor binding sites, *Nucleic Acids Research*, 31(13):3580-3585.

[43] Tompa, M. et al. (2005) Assessing computational tools for discovery of transcription factor binding sites, *Nature Biotechnology*, 23(1):137-144.

[44] Wagner, A. (1998) Distribution of transcription factor binding sites in the yeast genome shows abundance of coordinately regulated genes, *Genomics*, 50:293-295.

[45] Wasserman, M. and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements, *Nature Reviews Genetics*, 5(4)276-287.

[46] Xie, X., Lu, J., Kulbokis, J., Golub, T., Mootha, V., Lindblad-Toh, K., Lander, E., and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals, *Nature*, 434:338-345.

[47] Yarragudi, A., Parfrey, L., and Morse, R. (2007) Genome-wide analysis of transcriptional dependence and probable target sites for Abf1 and Rap1 in Saccharomyces cerevisiae, *Nucleic Acids Research*, 35(1):193-202.

[48] Zhang, Z. and Dietrich, F. (2005) Mapping of transcription start sites in Saccharomyces cerevisiae using 5′ SAGE, *Nucleic Acids Research*, 33(9):2838-2851.

[49] Zhou, Q. and Liu, J. (2004) Modeling within-motif dependence for transcription factor binding site predictions, *Bioinformatics* 20:909916.