***CIS*: Compound Importance Sampling Method for Transcription Factor Binding Site p-value Estimation**

Yoseph Barash[1,*], Gal Elidan[1,*], Tommy Kaplan[1,2,*], and Nir Friedman[1,†]

[1]School of Computer Science & Engineering, The Hebrew University, Jerusalem, 91904 Israel

[2]Hadassah Medical School, The Hebrew University, Jerusalem, 91120 Israel

{hoan,galel,tommy,nir}@cs.huji.ac.il

Abstract

Transcriptional regulation is mainly obtained by transcription factors that bind sequence-specific binding sites, and control the expression of nearby genes. By modeling such sites one can scan regulatory regions, searching for putative binding sites of a specific factor, thus constructing a genome-wide regulatory network. Recently, several works demonstrated the importance of rich probabilistic models that capture inner-dependencies within binding sites. Here we present a general, accurate and efficient method for estimating the statistical significance of putative binding sites, applicable to any probabilistic binding site model. Finally, we demonstrate the accuracy of the method using synthetic and real-life data.

Keywords: Transcription Factor Binding Sites, Genome-wide Scans, p-value estimation, Importance Sampling.

1 Introduction

Detecting *cis*-regulatory motifs in long DNA sequences is a central problem in modern biology, as it offers a simple way for understanding the transcriptional regulation that control the expression of genes. Much effort has been put in gathering known transcription factor binding sites (TFBSs), and in finding probabilistic models for describing them [12]. Ideally, we would scan the genome using these models, find putative BSs (and target genes) of each factor, and construct a global regulatory network. This is usually done by scoring each sub-sequence in the promoter using the log of the ratio between its probability according to the binding model, and its probability according to the background model. These ratios (for a specific factors) are often refered to as a *position specific scoring matrix* (PSSM).

To identify putative BSs, we wish to estimate the statistical significance of each score, so we can eliminate all the subsequences whose scores are expected according to the null background model,

[*] These authors contributed equally to this manuscript

[†] Contact: Nir Friedman, Bauer Laboratory, 7 Divinity Ave. Cambridge MA 02138. phone: +1-617-953-9279

and consider only those with significant p-values. Yet, the task of *in silico* detecting BSs within regulatory regions has not been fully answered, mainly due to complications arising in eukaryotic genomes. Currently, the exact location of promoter regions is unknown for the vast majority of higher organisms, so we resort to searching motifs in non-coding regions, whose length often exceeds several thousands of bases. This results in a lower signal-to-noise ratio, and in many putative sites, caused simply by the immense number of subsequences we evaluate. For this, we need to correct the assigned p-values of each subsequence and account for multiple testing [5, 4, 9]. All these correction methods result in evaluating extremely small p-values, and so the need for accurate estimation is further emphasized.

Recently, several works [2, 10, 8] demonstrated the importance of modeling TFBSs using rich probabilistic models that allow for inner-dependencies within the positions of a binding site. Current methods for evaluating p-values are either only applicable to motif models that assume position independence (e.g. [1]) or rely on functional approximations of the scores' distributions, whose accuracy is somewhat arguable (e.g. [11]). In addition, due to the small magnitude of the p-values we're interested in, naïve approaches for p-value estimation, such as sampling sequences from the background model, and calculating the empirical distribution of scores, become impractical.

In this paper, we present a general, accurate and efficient method for estimating the statistical significance of putative binding sites, which is applicable to every choise of probabilistic binding model. Our *Compound Importance Sampling* (CIS) algorithm relies on the general concept of Importance Sampling [7], allowing to approximate a desired distribution by drawing weighted samples from another distribution. Thus, we are able to conceptually mimic the naïve sampling approximation, while using a much smaller sample set. CIS does not make any simplifying assumptions for either the motif or the background model, nor does it use a functional approximation for the scores' distribution. This enables us to apply it to any probabilistic form of the motif or the background model.

2 Computing the Statistical Significance

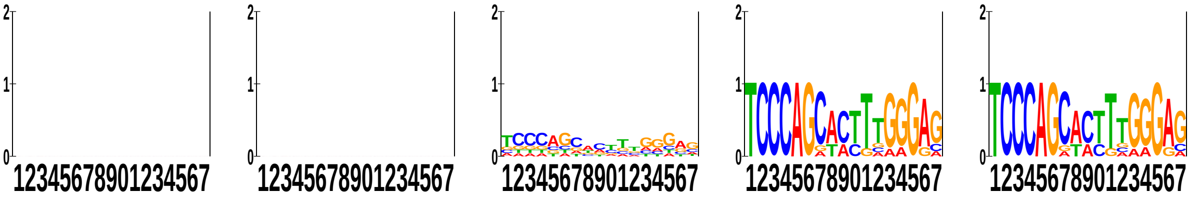
In a genome wide scan, each candidate binding subsequence X is given a score S. Estimating the statistical significance of the score amounts to estimating the probability of seeing a score that is as good or better when scanning sequences generated from some background distribution. That is

$$P - value(S) = \boldsymbol{E}_{P_{BG}(X)}[1\left\{Score(X) > S\right\}] \tag{1}$$

where $P_{BG}(X)$ is the background distribution and $1\{\}$ is the indicator function. The well founded log-odds score (e.g. [6]) of the probability of the subsequence the model $P_M(X)$ vs. the probability of the subsequence given the background $P_{BG}(X)$ is typically used in this scenario.

Analytic computation of the above p-value are often intractable despite the use of innovative methods such as the branch and bound method of [3] due to exponential dependence on the number of positions in the binding site model. When considering a general form of the motif and background the problem is even more complex. Analytic computations are usually not possible and highly effective methods that assume position independence (e.g. [1]) are not applicable.

Thus, we need to resort to some other approximation. One possibility is to assume some functional for the p-values distribution such as a normal distribution form (e.g. [11]). As we demonstrate in Section 4 this approach suffers from significant inaccuracies in the region of small p-



$Q_0 \equiv P_{BG}(X)$ Q_1 Q_2 Q_3 $Q_4 \equiv P_M(X)$

Figure 1: Illustration of the component of the proposal distribution $Q(X)$ when 5 components are used. The leftmost component is the background distribution, the rightmost component is the motif model distribution and in between are different level of mixtures between the two extremes.

values. Another possibility is to generate samples from the background distribution and use an empirical p-value estimation as a proxy to the true p-value. While in theory this approach is sound, it cannot be applied in practice for the case of genome wide scans. In this scenario we need to correct for multiple testing ([5, 4, 9]) and thus need to estimate in practice p-values in the order of 10^{-5} and lower. Even generating 10^6 samples leads to noisy estimation in the region of interest making the method impractical.

3 Compound Importance Sampling (CIS)

Our goal is to develop a general and effective method for estimating the statistical significance of binding sites in a genome wide scan. The discussion in Section 2 hints that we might want to use a sampling approach but somehow sample from more from the region of small p-values. In this section we present a method for doing so based on importance sampling.

We start with a short introduction of the method of *importance sampling*. Assume that we want to compute the expectation $\boldsymbol{E}_{P(X)}[f(X)]$ of some function $f(X)$ with respect to a distribution $P(X)$ and that we are unable to compute this using a closed form formula as in the case of p-value computation of Eq. (1). In addition, assume we are unable to generate samples from $P(X)$ directly so that we cannot use naive sampling in order to evaluate the desired expectation. Importance sampling (e.g. [7]) is a general method for evaluating the desired expectation in cases where we cannot sample from $P(X)$ but can still evaluate $P(x)$ for a particular sample x . The method uses a seemingly unrelated *proposal distribution* $Q(X)$ to generate samples and then evaluates the desired expectation using these samples and a correction weight. While this might sound surprising, it relies on the following simple equality

$$\boldsymbol{E}_{P(X)}[f(x)] = \int_x P(x)f(x)\frac{Q(x)}{Q(x)}dx = \int_x Q(x)\left[f(x)\frac{P(x)}{Q(x)}\right]dx = \boldsymbol{E}_{Q(x)}[f(x)w(x)]$$

where $w(x)$ is defined as the ratio $\frac{P(x)}{Q(x)}$. Thus, each sample from $Q(X)$ is re-weighted to account for the fact that it is not generated from the distribution of interest $P(X)$. [1]

While primarily used in cases where we cannot sample from $P(X)$, importance sampling is also valid when we simply prefer to sample from $Q(X)$ rather then $P(X)$. This observation allows

[1] The only requirement of $Q(X)$ is that is does not eliminate the possibility (gives zero probability) of any event x that has positive probability $P(X)$.

us to adapt the method of importance sampling for our needs. Recall, that we wanted to generate more samples from the region of interest in order to get precise estimation of low p-values. Thus, all we have to do is to define $Q(X)$ that will put more emphasis on the region of statistically significant scores and use

$$w(X) = \frac{P_{BG}(X)}{Q(X)}$$

when estimating the p-value of a putative binding site. To define an effective $Q(X)$, we first need to characterize the distribution that governs the region of interest. This can be done by considering samples (K-mers) that are generated from the motif model at hand. Since these samples are generated from the motif itself, their probability given the model is typically high and thus they receive significant scores. Accordingly, the region of low p-values is that of the distribution $P_M(X)$ of the motif model.

The above discussion might lead to the conclusion that we should simply set $Q(X) \equiv P_M(X)$. However, recall that we want to compute p-values with respect to the background distribution $P(X) \equiv P_{BG}(X)$. Sampling only from $P_M(X)$ completely ignores the distribution of interest. This will typically results in assigning extremely small weight to the samples generated from Q and will again result in a poor overall estimation of the p-value with respect to $P_{BG}(X)$.

This suggest that we want to sample both from $P_M(X)$ and from $P_{BG}(X)$, say M_1 and M_2 samples respectively. This is equivalent to defining $Q(X) \equiv \frac{M_1}{M_1+M_2} P_{BG}(X) + \frac{M_2}{M_1+M_2} P_M(X)$ and compute the appropriate weights for each of the M samples. This approach takes into account both extreme distributions. One extreme will provide the general form of $P_{BG}(X)$ while the other will concentrate on the region of interest characterized by $P_M(X)$.

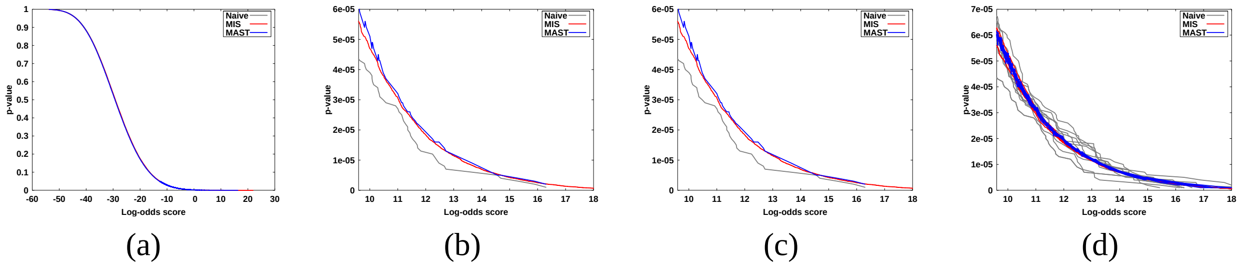
A possible caveat of the above approach is that by taking into account only the two extreme distributions, our p-value estimation of "middle-ground" scores will not be accurate. [2] Thus, we want to refine the approach in order to consider a combination of distributions that are a smoothed mixture of these two extremes. We define the *Compund Importance Sampling* (CIS) approach for estimating p-values of putative binding sites as follows: We L distribution that are a smoothed mixture between the background and the motif model. That is the proposal distribution is

$$Q(K - mer) = \sum_{i=0}^{L} m(Q_i) Q_i(K - mer)$$

where Q_0 is the background model, Q_L is the motif model and $m(Q_i)$ is the fraction of samples generated from the model. We can now use this mixture $Q(X)$ and use importance sampling as is to evaluate the desired p-values.

To illustrate the concept of CIS, Figure 1 shows an example of a proposal distribution with 5 component models. For each model we show the sequence-logo where for each position, the height of a letter is proportional to the log-probability of seeing that letter in the specific position. On the left is the distribution generated from the background model and on the right the one of the motif model. It is important to emphasize that sampling from Q_3, for example, is not equivalent to a weighted sum of samples from $P_{BG}(X)$ and $P_M(X)$. In fact, a sample typically generated from Q_3 is very unlikely to be generated from either of the extremes.

[2] This was indeed confirmed in preliminary experiments not presented here for lack of space



(a) (b) (c) (d)

Figure 2: Comparison of p-value estimation for the RAP1 motif model from the TRANSFAC database [12] using a 3-order Markov model of *S. cerevisiae*. Shown is the **Naive** sampling approximation of the true p-values using 10^6 samples, the **Normal** approximation to the p-value distribution, the method used by **MAST** and our **CIS** importance sampling method that use 40,822 samples. Shown is the cummulative p-value (y-axis) vs. the log-odds (x-axis). (a) the case of a PSSM model where positions are independent (b) zooms in on the area of p-values below 10^{-3} (c) the case of a Mixture of Trees of model where **MAST** is not applicable (d) shows 10 repeats of the two sampling methods **Naive** and **CIS** FIX THIS WITH UPDATED FIGURES

It is important to note that our *Compund Importance Sampling* (CIS) approach makes no assumptions on the form of the distributions and can thus take into account any joint distribution of the motif positions $P_M(X)$ or the background distribution $P_{BG}(X)$. In fact, no assumption was made also on the form of the score making the method applicable to very general settings of a genome scan. As we demonstrate in Section 4 the CIS method evaluates accurate p-values and requires an order of magnitude less samples then the naive sampling approach.

4 Experimental Validation

We want to demonstrate the effectiveness of our *Compund Importance Sampling* (CIS) method described in Section 3 in computing the statistical significance p-value for putative binding sites in a genome wide scan. As a sanity check, we start by comparing the different methods discussed in Section 2 to CIS in the case of a simple *Position Specific Scoring Matrix* (PSSM) motif model that assumes position independence and a 3-order Markov background model. For our CIS algorithm we use a mixture of 10 motif models. We sample 10,000 samples from the $Q_0(K - mer)$ background model and span the range to 1,000 samples for the $Q_L(K - mer)$ motif model totalling 40,822 samples.

We use the 14-Mer model of the RAP1 transcription factor from the TRANSFAC [12] database. We generate 10^6 14-Mers from the 3-order Markov background model of *S. cerevisiae*. This allows us to evaluate reasonable empirical values in the order of 10^{-5}. For each 14-Mer, we compute the log-odds score of its probability given the RAP1 model and the background model. We then use each of the following method to assign a statistical p-value for each of the 10^6 scores: our **CIS** algorithm, **Normal** functional approximation as in [11] and the method used by **MAST** [1]. Figure 2(a) shows the cummulative distribution of p-values of each method vs. the log-odds score in the region of p-values lower then 10^{-3}. It is clear that both **CIS** and **MAST** are superior to the **Normal** approximation method and follow the naive sampling method similarly with a slight advantage for **CIS**.

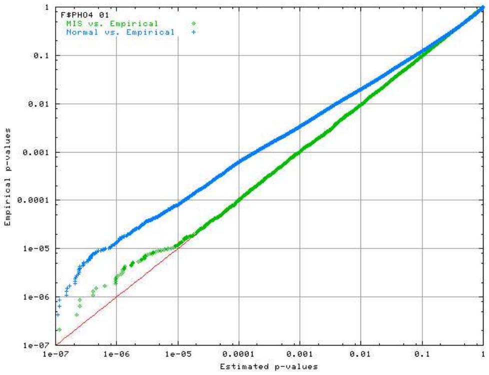


Figure 3: NEED TO COMPLETE THIS

We now turn to the case where there are dependencies within the positions in the motif model and the extremely effective **MAST** method can no longer be used. Specifically, we use a mixture of trees model (see [2]). Figure 2(b) shows the evident superiority of **CIS** when compared to the **Normal** approximation. While **CIS** follows the *Naive* sampling method reasonably, the deviation between the two methods is still a cause for concern. Specifically, since the naive method uses 10^6 samples while **CIS** uses only 40,822 samples, one might suspect that the **CIS** suffers from robustness problem to the sampling process.

To evaluate this explicitly, we repeated the experiment 10 times for both sampling methods. Figure 2(c) compares the range of these runs. It is evident that **CIS** shows strong robustness when compared with the naive method. In fact, this results shows that **CIS** can follow even a naive sample of 10^7 effectively using 0.4% the number of samples.

We now want to demonstrate the importance of the CIS method in a true genome scan. A crucial factor in such a scan is the amount of false positive signals that are considered for a given level of true positive rate. To evaluate this, we estimate the signal-to-noise ratio (true positive to false positive) when scanning the promoter regions of *S. cerevisiae* with the TRANSFAC motif KAKA_YPD (see http:www.OURSITE) for results for a range of TRANSFAC motifs). Figure 3 shows... COMPLETE THIS WHEN WE HAVE FIGURE

5 Discussion

In this work we introduced a general and efficient method for estimating the statistical significance of putative binding site in a genome wide scan. We demonstrated the accuracy of the method when using simple as well as models that allow complex dependencies within the binding positions on real-life like synthetic experiments and actual scan of promoter regions.

To our knowledge, this is the first method for evaluating statistical significance of putative binding sites that can be applied to general forms of the motif model, the background model and the scoring function. Thus, the method can be extended to other setting beyond the scope of transcription regulation such as splicing event mechanism and ??? (COMPLETE THIS) Another interesting challenge is to extend our method for the case of motif complexes that include several general motif models

References

[1] T. L. Bailey and M. Gribskov. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14:48–54, 1998.

[2] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan. Modeling dependencies in Protein-DNA binding sites. *Proc. Seventh Inter. Conf. Res. in Comp. Mol. Bio. (RECOMB)*, 2003.

[3] G. Bejerano. Efficient exact value computation and applications to biosequence analysis. In *Proceedings of the seventh annual international conference on Computational molecular biology*, pages 38–47. ACM Press, 2003.

[4] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *J. Royal Statistical Society B*, 57:289–300, 1995.

[5] C. E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60, 1935.

[6] M. H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

[7] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.

[8] H. Huang, M.J. Kao, X. Zhou, J.S. Liu, and W.H. Wong. Determination of local statistical significance of patterns in markov sequences with application to promoter element identification. *Journal of Computational Biology*, pages , to appear, 2004.

[9] Storey JD and Tibshirani R. Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, 100:9440–9445, 1995.

[10] O. D. King and F. P. Roth. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res*, 31(19):e116, 2003.

[11] S. Sinha and M. Tompa. A statistical method for finding transcription factor binding sites. In *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, volume 8, pages 344–54. 2000.

[12] E. Wingender, X. Chen, Fricke E., R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhauser, M. Pruss, F. Schacherer, S. Thiele, and S. Urbach. The TRANSFAC system on gene expression regulation. *Nuc. Acids Res.*, 29:281–283, 2001.