BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 5 2005, pages 596–600
doi:10.1093/bioinformatics/bti041

Sequence analysis

CIS: compound importance sampling method for protein–DNA binding site *p*-value estimation

Y. Barash[1,†], G. Elidan[1,†], T. Kaplan[1,2,†] and N. Friedman[1,*]

[1]School of Computer Science & Engineering, The Hebrew University, Jerusalem 91904, Israel and
[2]Hadassah Medical School, The Hebrew University, Jerusalem 91120, Israel

Received on April 11, 2004; revised and accepted on September 14, 2004
Advance Access publication September 28, 2004

ABSTRACT

Motivation: A key aspect of transcriptional regulation is the binding of transcription factors to sequence-specific binding sites that allow them to modulate the expression of nearby genes. Given models of such binding sites, one can scan regulatory regions for putative binding sites and construct a genome-wide regulatory network. In such genome-wide scans, it is crucial to control the amount of false positive predictions. Recently, several works demonstrated the benefits of modeling dependencies between positions within the binding site. Yet, computing the statistical significance of putative binding sites in this scenario remains a challenge.

Results: We present a general, accurate and efficient method for computing *p*-values of putative binding sites that is applicable to a large class of probabilistic binding site and background models. We demonstrate the accuracy of the method on synthetic and real-life data.

Availability: The procedure for scanning DNA sequences and computing the statistical significance of putative binding site scores is available upon request at http://compbio.cs.huji.ac.il/CIS/

Contact: nir@cs.huji.ac.il

INTRODUCTION

Accurate detection of *cis*-regulatory elements in long DNA sequences is a central challenge in modern biology, as it offers a direct way for understanding transcriptional regulation and the expression of genes. Accordingly, extensive efforts have been put in gathering known transcription factor binding sites (Wingender *et al.*, 2001), and in finding models that characterize them. These models facilitate a systematic scan of genomic sequences to identify transcription factor target genes. In this article, we are interested in the following task: assuming that we have a model that characterizes the binding preferences of a particular transcription factor, we want to perform a genome-wide scan for its binding sites. A fundamental challenge in performing such a scan is in controlling the number of prediction errors. This problem is further emphasized in eukaryotic genomes where binding sites appear in extremely long intergenic regions. As a consequence, there is high probability of finding spurious binding sites due to the immense number of putative sites evaluated. To control the amount of this false positive noise in our predictions, we assign each possible site a score, and estimate its statistical significance. That is we compute how likely it is to find a score that is at least as good by chance.

Formally, the *p*-value of a putative binding site with score S is the probability of achieving this or higher score according to the background distribution over sites. This can formally be written as

$$\boldsymbol{E}_{P_{\mathrm{BG}}(X)}[1\{\mathrm{Score}(X) \geq S\}] \quad (1)$$

where $1\{\ \}$ is the indicator function, $P_{\mathrm{BG}}(X)$ is the background distribution over a random variable X that ranges over possible DNA motifs, and Score() is the scoring function. In a probabilistic framework, X is usually a subsequence of a fixed length and the score of each subsequence is the log of the ratio between its probability according to the binding site model, and its probability according to a background model (log-odds score).

The formulation of Equation (1) suggests a simple procedure for estimating the *p*-value of a score S in which we sample i.i.d. samples from $P_{\mathrm{BG}}(X)$, and then compute the fraction of samples whose scores are as high as S. Such a naive sampling procedure can reliably estimate *p*-values that are at least two orders of magnitude larger than the inverse of the number of samples. Thus, to estimate a *p*-value in the order of 10^{-3} we need about 10^5 samples.

It might seem that *p*-values from the magnitude of 10^{-3} are sufficient for finding statistically significant binding sites. Recall, however, that the typical application involves scanning long sequences. In this scenario, we treat each subsequence as a putative binding site, and then correct for multiple testing (e.g. Benjamini and Hochberg , 1995).[1] Assuming a typical promoter length of size 500 bp, these corrections result in the need for estimating *p*-values in the order of 10^{-5} or even lower, rendering the naive sampling approach impractical as it requires the order of 10^7 samples.

More sophisticated approaches either derive efficient analytical algorithms for the exact *p*-value (Wu *et al.*, 2000; Huang *et al.*, 2004), or use large deviation approximations (Bailey and Gribskov, 1998). These approaches are effective for probabilistic profile models (also known as PSSMs or PWMs) that assume that the probability of nucleotides at one position of the binding site is independent of all other positions.

*To whom correspondence should be addressed.

†The authors wish to be known, that in their opinion, the first three authors should be regarded as joint First Authors.

[1]An alternative model is to compute the *p*-value for the best score of a subsequence within a long sequence. Such a model explicitly deals with the dependencies between the putative binding sites. However, the computations of *p*-values in this model introduce several technical problems. In practice, treating the evaluation of each binding site as an individual hypothesis test and then correcting for multiple tests does not introduce noticeable bias.

Recently, several works demonstrated the importance of modeling transcription factor binding sites using probabilistic models that allow for inner-dependencies within the positions of a binding site (Barash *et al*., 2003; King and Roth, 2003; Zhou and Liu, 2004). Such models provide richer representations of the binding preferences of a transcription factor, and consequently can better discriminate sites that match these preferences. This can lead to more accurate binding site identification. Yet, the question of assigning p-values for putative binding sites when using dependency models still remains open. Specifically, analytical methods that are designed for PWMs are not applicable for these richer models.

In this article, we present a general, accurate and efficient method for estimating p-values. Our compound importance sampling (CIS) algorithm uses importance sampling (Hammersley and Handscomb, 1964), and allows us to conceptually mimic the naive sampling approximation using a significantly smaller number of samples.

As we describe below, CIS is applicable for a wide range of binding site models. Specifically, our implementation of CIS is based on Bayesian networks models. This covers many commonly used models as a special case, such as Markov models, trees and mixture of PSSMs. We demonstrate the accuracy and efficiency of the CIS method on synthetic and real-life data for the case of simple position-independent models as well as for models that allow dependencies, where standard methods cannot be applied.

METHODS

To estimate the p-value of the score S of a candidate site, one needs to estimate Equation (1). In the naive sampling approach this is carried out by sampling from the background distribution $[P_{\mathrm{BG}}(X)]$ and computing the fraction of samples where $\mathrm{Score}(X) \geq S$. As discussed in the previous section, sampling directly from P_{BG} requires an unreasonable number of samples to reliably estimate small p-values, since for most samples $\mathrm{Score}(X) < S$. This suggests that it would be better to try and sample assignments of X that have higher scores. In doing so, we have to make sure that we still compute correct p-values of these scores with respect to P_{BG}.

The compound importance sampling approach

Importance sampling (Hammersley and Handscomb, 1964) is a general method that estimates $\boldsymbol{E}_{P(X)}[f(X)]$, where $f(X)$ is some function over X, using samples from a proposal distribution $Q(X)$. It is especially useful in cases where sampling directly from $P(X)$ is not possible. The method relies on the following equalities

$$\begin{aligned}\boldsymbol{E}_{P(X)}[f(X)] &= \sum_x P(x)f(x) = \sum_x P(x)f(x)\frac{Q(x)}{Q(x)} \\ &= \sum_x Q(x)\left[f(x)\frac{P(x)}{Q(x)}\right] \\ &= \boldsymbol{E}_{Q(X)}[f(X)w(X)], \end{aligned} \tag{2}$$

where the weight $w(x) \equiv P(x)/Q(x)$ compensates for the bias introduced by sampling from $Q(X)$ rather than $P(X)$. Estimating $\boldsymbol{E}_{P(X)}[f(X)]$ using samples from $Q(X)$ is therefore basically the same procedure as naive sampling, but here each sample x is re-weighted according to $w(x)$. The total weight of the samples in this case is not the number of samples N but rather $\sum_i^N w(x_i)$.

To apply this general framework to estimate p-values, we use the formulation of Equation (1), where $f(x) = 1\{\mathrm{Score}(x) \geq S\}$. The main decision in applying importance sampling is the choice of the proposal distribution Q. For the log-odds score, values of X sampled from the binding site model are more likely to receive high scores. Naively, we can set Q to be the distribution of the binding site model $P_M(X)$, and directly sample from the region of high scores. This approach, however, is problematic since we need a fair amount of low scoring samples, even when estimating the p-values of high scoring samples. Without such samples the empirical histogram of observed scores would be biased towards high scores, and the empirical estimation of the p-values would be poor (data not shown).

One possible solution is to sample a mixture of n_1 samples from the model distribution P_M and n_2 samples from the background P_{BG}. This is equivalent to sampling from

$$Q(X) = \frac{n_1}{n_1+n_2}P_M(X) + \frac{n_2}{n_1+n_2}P_{\mathrm{BG}}(X) \tag{3}$$

While this solution takes into account both extremes, in practice it still suffers from poor estimation of the 'middle-ground' scores. Thus, we refine the above approach and consider a combination of richer set of models. We define CIS as:

$$Q(X) = \sum_{\alpha\in\mathcal{A}} n_\alpha Q_\alpha(X) \tag{4}$$

where $\mathcal{A}$ is an index set and n_α is the fraction of samples generated from the model Q_α. The models $\{Q_\alpha : \alpha \in \mathcal{A}\}$ are basically 'smoothed' versions of $P_M(X)$ that bias it in different degrees toward $P_{\mathrm{BG}}(X)$.

Characterizing *CIS*'s mixture distributions

The main question we now face is how to define the set of mixture component's $\{Q_\alpha\}$ that will effectively 'smooth' $P_M(X)$ towards $P_{\mathrm{BG}}(X)$. Specifically, we want to create distributions $\{Q_\alpha\}$ that will enable us to use a relatively small number of samples and, at the same time, provide a good approximation for the distribution of $\mathrm{Score}(X)$ over the entire range of scores.

The method we suggest is based on the following formulation: let $X = X_1, \ldots, X_k$ denote the k binding site's positions. Using the basic chain rule for any multivariable distribution we can write:

$$P(X) = \prod_{i=1}^{k} P(X_i \mid X_1, \ldots, X_{i-1}) \tag{5}$$

where P is either $P_{\mathrm{BG}}(X)$ or $P_M(X)$. We can now define Q_α as:

$$\begin{aligned}Q_\alpha(X) = \prod_{i=1}^{k}[&\alpha P_M(X_i \mid X_1, \ldots, X_{i-1}) \\ &+ (1-\alpha)P_{\mathrm{BG}}(X_i \mid X_1, \ldots, X_{i-1})]\end{aligned} \tag{6}$$

It can easily be shown that Q_α is indeed a proper distribution.

The basic difference in the above formulation from Equation (3) is that it mixes the conditional probability of each position in the binding site separately. As an example, consider a simple case in which the the background distribution is a zero-order Markov model favoring Guanine (G) with high probability for each position, and the binding site model is a 5 bp long PSSM favoring Adenine (A) with high probability for each position. In this case sampling from a distribution defined as Equation (3) would result in many high scoring AAAAA as well as many low scoring GGGGG samples. 'Mid range' samples, such as GAGAG would be very rare, degrading the evaluation of the overall p-value range. When using Equation (6) with $\alpha = 0.5$, on the other hand, GAGAG is as likely as the two extremes when assuming that positions are independent of each other.

The idea of CIS is further illustrated in Figure 1. Figure 1a shows the sequence logo of the background (Q_0) and the binding site (Q_1) defining the extreme distributions as well as one of the middle ground mixture distributions ($Q_{0.5}$). The effect of sampling from each of these component is illustrated in Figure 1b, where the corresponding score distributions are shown. Sampling only from $P_M(X)$ (corresponding to Q_1 in Figure 1) results mostly in high scoring samples. This is problematic since the total weight of the scores

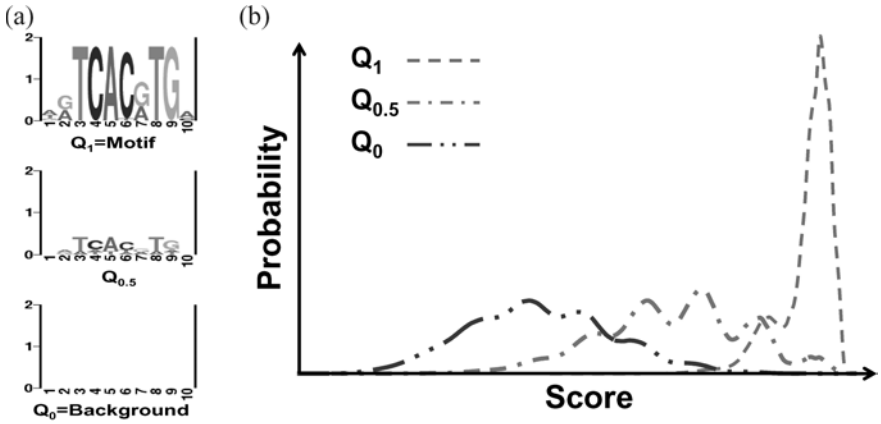

Fig. 1. Illustration of a proposal distribution $Q(X)$. (**a**) Sequence logos for several components of the porposal distribution, ranging from the binding site model (top), to the background model (bottom) where the information content of each position is low. (**b**) The distributions of log-odd scores for the three models.

accumulates extremely slowly resulting in a poor overall p-value estimation. If, on the other hand, we sample only from $P_{\mathrm{BG}}(X)$ (Q_0 in Figure 1), we get the right distribution but require a large number of samples for accurate estimation of the p-values of high scores. Incorporating samples from a mixture distribution as $Q_{0.5}$ in Figure 1 offers a better coverage of the middle ground. As we demonstrate in the next section, combining a whole set of such mixture distributions results in improved p-value estimation for the full range.

We emphasize that the formulation is not limited to any specific form of distribution. The only assumption we make in Equation (5) about $P_M(X)$ or $P_{\mathrm{BG}}(X)$ is that we can compute the conditional probability of X_i, given a specific value assignment to the preceding variables. In using Q_α, we either sample an instance or compute the probability of a given instance. In both cases, these are the only queries we need. Also note that we choose the order of variables in which to perform this expansion. In some cases, a specific order (e.g. one that conforms to the topological order in a Bayesian network) can lead to more efficient computations.

In the case of Bayesian networks, Equation (5) and consequently Equation (6), decompose according to the dependency model, where each X_i is dependent on a (typically small) number of additional positions. In the case of PWMs, for example, each position is independent of all other positions, while in K-order Markov model each position depends only on K others. Therefore, using Equation (6) rather than the naive mixture of Equation (3) results in an inherently different set of samples. Similarly to the above illustrative example, this 'smoothed' mixing results in better coverage of the middle ground score range and leads to an overall accurate estimation of p-values.

Finally, when using CIS we have to decide on the number of components as well as the number of samples and degree of smoothing (α) for each component. In this work, we use 10 components for which the α coefficient vary linearly from 0 to 1. The number of samples drawn from each component decreases exponentially, starting from 10 000 samples from $P_{\mathrm{BG}}(X)$ to 1000 samples from $P_M(X)$. It should be noted that the CIS method was proved to be robust in a wide range of settings.

RESULTS

As a case study, we examine the binding site model of RAP1 in *Saccharomyces cerevisiae* from TRANSFAC 7.3 (Wingender *et al*., 2001), which is 14 bp long. We estimated the p-value of each score using the following methods: CIS algorithm using 40 000 samples from a proposal distribution as illustrated in Figure 1; MAST (Bailey and Gribskov, 1998); functional approximation by normal distribution, where we estimate the mean and variance of Score(X) according to $P_{\mathrm{BG}}(X)$, and then use the tail probability of normal distribution as the p-value estimate.

As a proxy to the truth, we computed the p-values using the naive sampling procedure with 10^9 samples as described above. Figure 2a compares the p-value estimates by the different methods. While all methods appear the same, zooming into the region of interest in Figure 2b reveals significant discrepancies. It is evident that the normal approximation is inaccurate in this region. Both CIS and MAST provide accurate estimations, with a slight advantage to the CIS method.

Figure 2c shows evaluation of p-values for a binding site model with dependencies between positions studied by Barash *et al*. (2003) for the PHO4 transcription factor. For models such as this, MAST is not applicable. As we can see, the estimations of CIS are similar to the direct sample estimate. One might suspect that the slight deviation observed between the two curves is due to the smaller number of samples used by CIS. However, when comparing 10 repetitions of each procedure shown in Figure 2d, we see that CIS estimates are more robust than the ones by naive sampling while using two orders of magnitude fewer samples.

The robustness of the CIS estimator is furthered illustrated in Figure 3. Here the relative error in p-value estimation is plotted as a function of the estimated p-value (a), and of the sample size for two fixed p-values (b). In both graphs the advantage of using CIS with only 40 000 samples over using naive sampling with 10^6 samples is clear, particularly for p-values lower than 10^{-3}. As we can see, CIS with 40 000 samples provides a good compromise between efficiency and accuracy.

So far we demonstrated the effectiveness of CIS with respect to the background distribution directly. We conclude by demonstrating our approach on a real-life genome-wide scan. Using the Chromatin immunoprecipitation location analyses of Lee *et al*. (2002), we excluded all the promoter regions of *S.cerevisiae* genes that were found to be targets of the ZAP1 transcription factor. We then used the dependency model for ZAP1's binding sites by Barash *et al*. (2003) to scan the promoter regions of remaining genes of *S.cerevisiae*. Given that we removed ZAP1 targets, we expect that the remaining promoters will contain only few real binding sites of ZAP1. We used a third-order Markov model as a background distribution, and plotted the empirical frequencies of high scoring subsequence.

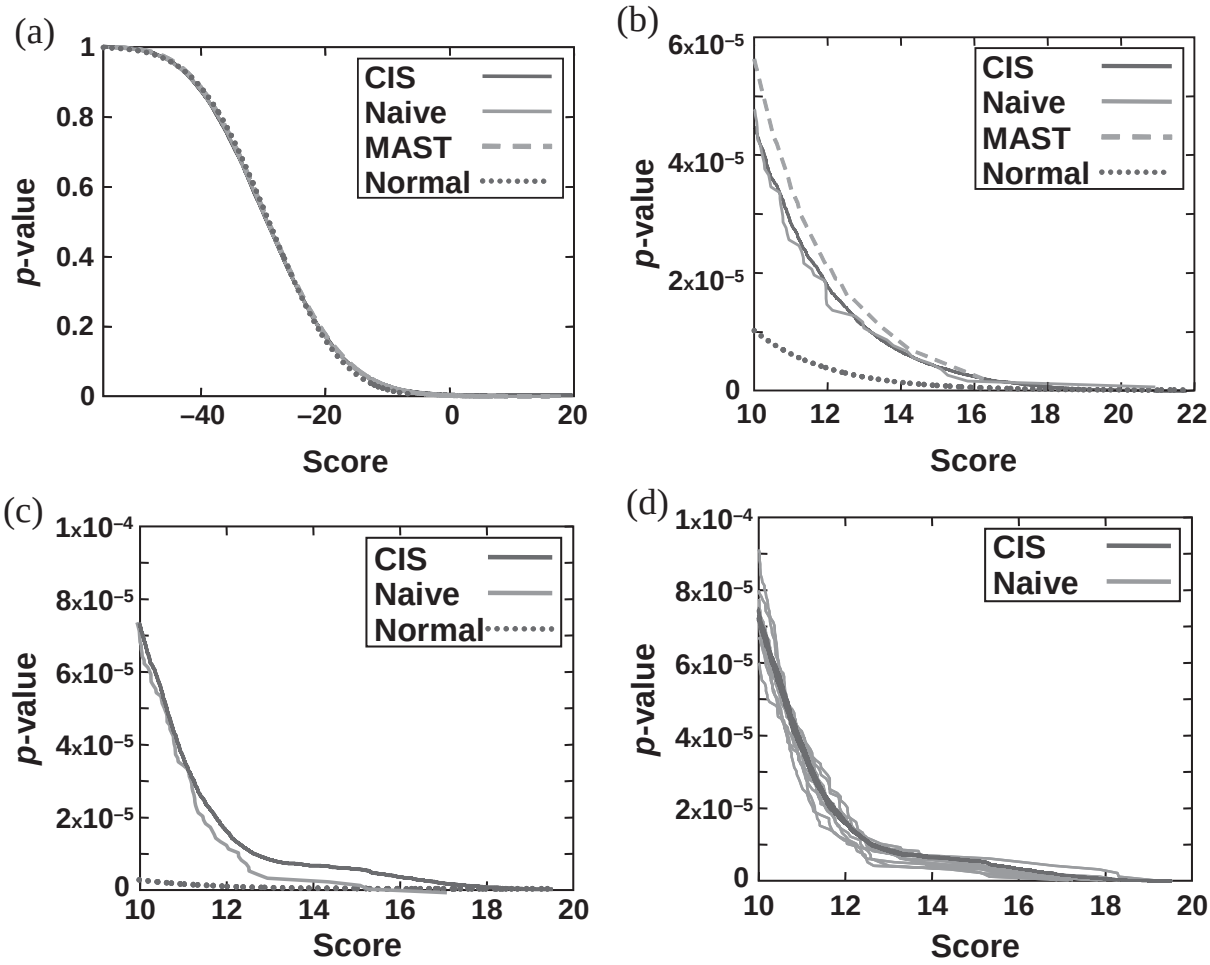

Fig. 2. *p*-value estimations for putative binding sites. Compared are the naive sampling approach using 10^6 samples, the normal approximation, MAST, and our CIS method (with 40 000 samples). Shown is the *p*-value (*y*-axis) as a function of the log-odds score (*x*-axis). (**a**) For the position independent model of TRANSFAC's RAP1; (**b**) Same as (a), zoomed on *p*-values $< 10^{-4}$; (**c**) For the dependency model of PHO4 studied by Barash *et al.* (2003) (here MAST is not applicable); (**d**) Test for robustness with 10 repeats of (c).

We note that in such a 'real-life' test, the accuracy of the results can be affected by factors other than the sampling technique used. These include the accuracy of the binding site and background models, and the fact we use i.i.d. subsequence samples instead of long sequences. Figure 4 shows that once again the normal approximation results in poor *p*-value estimates. More importantly, using the models described above and 40 000 i.i.d. samples, the CIS method provides accurate estimations over a wide range of *p*-values.

DISCUSSION

In this work we introduced a general and efficient method for estimating the statistical significance of putative binding sites in genome-wide scans. We demonstrated the accuracy of the method on both synthetic and genomic data, using simple as well as rich probabilistic models.

The CIS algorithm offers a practical method to handle a wide range of probabilistic representations of binding sites and background models. At the theoretical level, the only formal constraint of the CIS algorithm is for probability distributions that can be easily sampled from, or efficiently used to compute the conditional probability of any subsequence. This can be done in exact form in graphical models, such as Bayesian and Markov networks with small tree width (Pearl, 1988; Jensen, 1996). In fact, our implementation is based on Bayesian network representation of the models. For models where these queries are infeasible, one can adopt our methods to work with approximate inference techniques. This issue remains open for future research.

The general framework of the CIS algorithm makes it applicable to other tasks that involve the identification of sequence motifs, such as the identification of splicing junctions, the detection of protein motifs, etc.

In a broader theoretical context, *p*-value estimation can be viewed as an instance of the well-studied statistical problem of estimating the ratio between two normalizing constants (Meng and Wong, 1996; Chen and Shao, 1997; Gelman and Meng, 1998). The statistical literature for this problem includes a range of approaches that may be applicable to our problem. Several of these approaches are based on using importance sampling with different choices of sampling distributions [see Chen and Shao (1997) for a review]. One case that

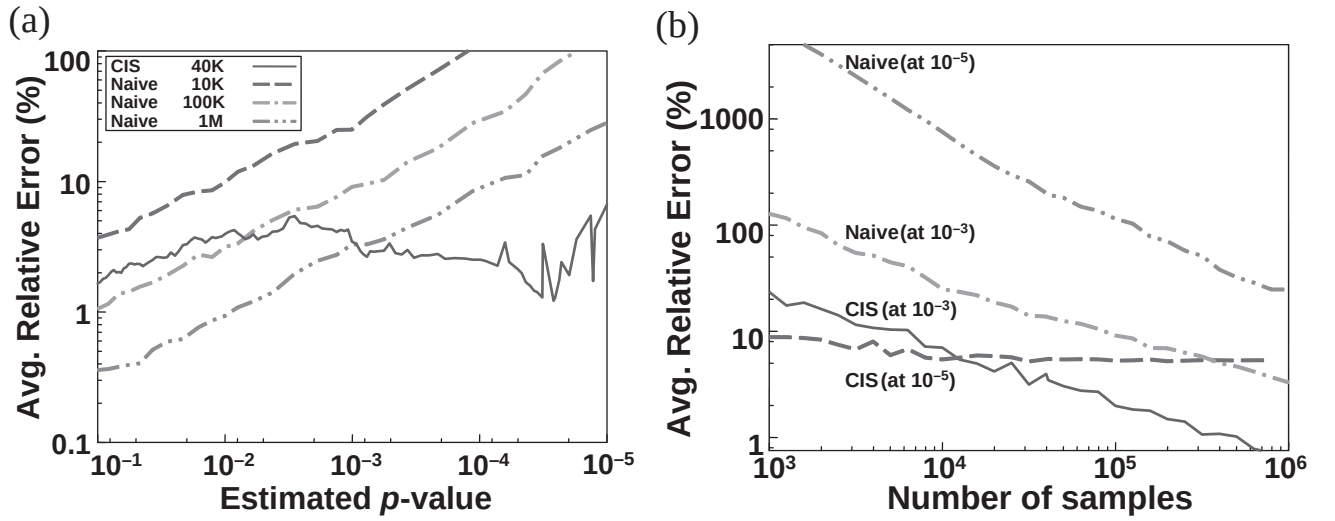

Fig. 3. Robustness of p-value estimation. Samples (10^9) from a third-order Markov background model were used to approximate the scores' p-values to the order of 10^{-5}. The biding site model is TRANSFAC's F_CBF1_B. For each setting 50 runs were made to compute the average relative error in p-value estimation (**a**) as a function of the estimated p-value and (**b**) as a function of the sample size at p-values 10^{-3} and 10^{-5}.

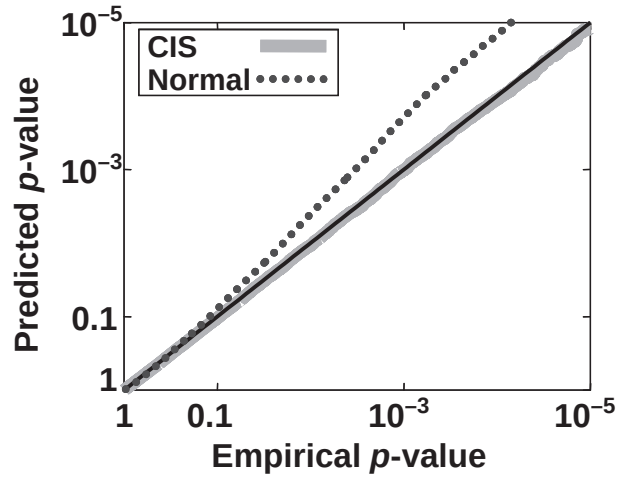

Fig. 4. p-value estimation for a genome-wide scan of *S.cerevisiae* using the ZAP1 dependency model studied by Barash *et al*. (2003) with a third-order Markov background model. Shown are the normal approximation of the p-values and the CIS estimates compared to the empirical frequency of these scores.

has received some analytical treatment is the proposal distribution of Equation (3) (Chen and Shao, 1997). However, to the best of our knowledge, there is no analysis of the mixture distribution we use here. This raises the question of the theoretical properties of our estimator and its applicability in a wider context. These questions, however, are beyond the scope of this work.

ACKNOWLEDGEMENTS

We thank Noa Shefi and the anonymous reviewers for useful comments on an earlier version of this manuscript. This work was supported in part by the Israel Science Foundation (ISF), and the Israeli Ministry of Science. Y. Barash was supported by an Eshkol fellowship. G. Elidan and T. Kaplan were supported by Horowitz fellowships. N. Friedman was supported by an Alon fellowship and the Harry & Abe Sherman Senior Lectureship in Computer Science.

REFERENCES

Bailey,T.L. and Gribskov,Y. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.

Barash,Y, Kaplan,T., Friedman,N. and Elidan,G. (2003) Modeling dependencies in protein-DNA binding sites. *Proceedings of the 7th International Conference on Research in Computational Molecular Biology (RECOMB)*, Berlin, pp. 28–37.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.

Chen,M.H. and Shao,Q.M. (1997) On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.*, **25**, 1563–1594.

Gelman,A and Meng,X.L. (1998) Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, **13**, 163–185.

Hammersley,J.M. and Handscomb,D.C. (1964) *Monte Carlo Methods*. Wiley, New York.

Huang,H. and Kao,M.J. and Zhou,X. and Liu,J.S. and Wing,W.H. (2004) Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification. *J. Comput. Biol.*, **11**, 1–14.

Jensen,F.V. (1996) *An Introduction to Bayesian Networks*. University College London Press, London.

King,O.D. and Roth,F.P. (2003) A non-parametric model for transcription factor binding sites. *Nucleic Acids Res.*, **3119**, e116.

Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Meng,X.L. and Wong,W.H. (1996) Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat. Sinica*, **6**, 831–860.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, CF, San Francisco, CA, USA.

Wingender,E., Chen,X, Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhausser,R. *et al*. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.

Wu,T.D., Nevill-Manning,C.G. and Brutlag,D.L. (2000) Fast probabilistic analysis of sequence function using scoring matrices. *Bioinformatics*, **16**, 233–244.

Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.