The Limits of De Novo DNA Motif Discovery

David Simcha[1,*], Nathan D. Price[2], Donald Geman[3]
1 Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, U.S.
2 Institute for Systems Biology, Seattle, WA, U.S.
3 Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, U.S.
∗ **E-mail: dsimcha1@jhmi.edu**

Abstract

A major challenge in molecular biology is reverse-engineering the cis-regulatory logic that plays a major role in the control of gene expression. This program includes searching through DNA sequences to identify "motifs" that serve as the binding sites for transcription factors or, more generally, are predictive of gene expression across cellular conditions. Several approaches have been proposed for *de novo* motif discovery – searching sequences without prior knowledge of binding sites or nucleotide patterns. However, unbiased validation is not straightforward.

We consider two approaches to unbiased validation of discovered motifs: testing the statistical significance of a motif using a DNA "background" sequence model to represent the null hypothesis and measuring performance in predicting membership in gene clusters. We demonstrate that the background models typically used are "too null," resulting in overly optimistic assessments of significance, and argue that performance in predicting TF binding or expression patterns from DNA motifs should be assessed by held-out data, as in predictive learning. Applying this criterion to common motif discovery methods resulted in universally poor performance, although there is a marked improvement when motifs are statistically significant against real background sequences. Moreover, on synthetic data where "ground truth" is known, discriminative performance of all algorithms is far below the theoretical upper bound, with pronounced "over-fitting" in training.

A key conclusion from this work is that the failure of *de novo* discovery approaches to accurately identify motifs is basically due to statistical intractability resulting from the fixed size of co-regulated gene clusters, and thus such failures do not necessarily provide evidence that unfound motifs are not active biologically. Consequently, the use of prior knowledge to enhance motif discovery is not just advantageous but necessary.

An implementation of the LR and ALR algorithms is available at http://code.google.com/p/likelihood-ratio-motifs/.

Author Summary

Introduction

A major undertaking in computational biology is to reverse-engineer the cis-regulatory logic of the transcriptome that underlies much of gene regulation. Cis-regulatory logic controls transcription on the same DNA molecule as the regulatory logic. In contrast, trans-regulatory logic affects transcription of sequences that may be on a different DNA molecule. Examples of cis-regulatory logic include transcription factor (TF) binding sites, and signals that affect nucleosome positioning [1], DNA melting [2] and DNA methylation [3].

Properties of DNA sequences which are predictive of the "expression profile" of a gene or a related cis-regulatory property such as TF binding are referred to as "motifs". Here an expression profile refers to the variation in the expression level of a gene across a variety of cellular conditions and the entire set of profiles may be quantized by assigning a gene membership in a cluster with other co-expressed genes;

prediction then refers to identifying gene clusters from motifs. Canonically, motifs represent transcription factor binding sites (TFBS), but could be any predictive feature. Motifs may be single sequences, sets of sequences or probability distributions over sequences or over sequence features. The presence of a motif may be defined categorically (e.g., present or absent) or quantitatively (e.g., with reference to a probability distribution). The regulatory role of motifs may be investigated individually or interactively. Our focus is the discovery of individual motifs rather than learning combinatorial logic after initial motif discovery, as addressed for example by Beer et al. [4]. In addition, we only consider *de novo* motif discovery, meaning the discovery of motifs from an unbiased search of a set of sequences, rather than knowledge-based motif discovery, for instance using previously cataloged TFBS.

Unbiased validation of motifs discovered by computational methods is not straightforward because neither the locations nor nucleotide pattern of functional cis-regulatory elements are known completely. There are two statistical approaches to validation, one based on discrimination and the other based on hypothesis testing.

The discriminative validation method is to treat motif discovery as a discrete classification problem. This can be done if DNA segments near genes are partitioned into disjoint clusters. The goal is then to find motifs that can discriminate among, and thereby characterize, the clusters. There are two principal ways of defining clusters. One is to quantize their expression profiles by applying a standard clustering algorithm such as K-Means to gene expression profiles; in this case, the expression profile corresponds to the gene adjacent to the sequence. The other is to aggregate sequences based on the presence or absence of a transcription factor binding to the sequence, as determined by ChIP-chip or ChIP-seq experiments, forming nearly disjoint clusters. A motif associated with a given cluster is then validated by measuring its performance in distinguishing that cluster from all others, with appropriate adjustments for overlap in the ChIP case.

Common unbiased methods for measuring performance are cross-validation and hold-out validation. Both methods avoid the over-optimism inherent in measuring accuracy by resubstitution, which is the use of the same data for training and validation. This over-optimism, known as "overfitting", is caused by the chance occurrence of patterns which are genuinely discrmininating in the training data but do not generalize to other data (e.g. held out samples) which are generated by the same mechanism or at least have the same statistical properties. We will show that these issues are pivotal in understanding the limits of *de novo* discovery. We use hold-out validation, which is both computationally simpler and generally more conservative than cross-validation. Data is partitioned into disjoint subsets and one is used for training and the other for validation. In our case, this means splitting each cluster of sequences (genes) into two groups, one used for motif discovery and the other for measuring classification performance. Notably, in the case of synthetic DNA sequence data with planted motifs, one can compute the Bayes error rate, or lowest error rate possible for a given classification problem, and related theoretical upper bounds on the performance of *de novo* methods.

Motifs can also be validated by statistical hypothesis testing. Properly done, this can improve interpretability by providing information on whether a motif is stronger than what one would expect to discover in a random set of DNA sequences. (The precise definitions of "stronger" and "random" vary depending on the method used, but the goal is conceptually the same.) Hypothesis testing is often neglected in the literature, especially when the motif model is probability-based (e.g., involves a position-weight matrix) rather than requiring an exact match to a pattern. When hypothesis testing is done, a generative null model is often used, which means defining a probability distribution over DNA sequences which represents "random" DNA not involved in any regulatory process. We argue (see Results) that the most common variants of these models are often "too null" . They fail to capture important properties of bulk, presumably non-regulatory, DNA. Especially in the human genome repetitive elements are prevalent. For example, repeats of the ALU element constitute approximately ten percent of the human genome [5]. Repetitive elements also have been shown to be highly conserved in some cases, with suggestive evidence of a regulatory role [6]. Therefore, masking such elements is questionable.

A second complicating factor in formulating a null hypothesis is the variation in low-order sequence properties such as dimer frequencies across clusters. Most methods do not consider these potentially discriminating signals; an exception is work that accounts for nucleosome positioning [7], which can be predicted by such low-order properties as the frequency of AA dimers. Assuming that low-order sequence properties are similar across clusters is one way that background models are "too null"; that is, they represent significantly less structure that exists in real biological sequences. Such properties appear (see Results) to be important predictors of expression profile and TF binding even by themselves. In addition to improving the biological relevance of hypothesis testing, using them as features might improve the performance of motif discovery algorithms.

These factors lead to motifs that are both statistically and biologically irrelevant being frequently declared significant. These kinds of discrepancies clearly reflect important differences between statistical significance (that varies relative to the chosen null hypothesis) and biological significance. Selecting better null distributions helps to align biological and statistical significance. In this article we argue that hypothesis testing based on discrimination is much more relevant than testing based on generative models. Here the null hypothesis is that a given motif is not overrepresented in a set of "foreground" sequences that share some property such as binding to a given TF or being upstream of a set of coregulated genes relative to "background" sequences, such as regulatory regions not adjacent to a coexpressed set of genes or not bound by the same TF as the foreground sequences. In this context repetitive elements are not problematic because, if a given element has no regulatory role, we expect it to occur in as large a fraction of the background sequences as foreground sequences.

In this paper, we explore the impact of background sequence models and variation of low-order sequence properties in validating existing motif discovery algorithms on both real and synthetic data, where we can compare the results obtained by *de novo* methods to the theoretical upper bound achieved when the correct motif is known. All methods are benchmarked using holdout validation, meaning evaluation of a classification rule using labeled data not used to train the classifier. To assess over-fitting, we also record the resubstitution (training set) error. We also introduce a new regression framework based on "mismatch" features (see Methods) designed to mitigate some of problems we have identified; adjusting for low-order sequence properties and an empirically-based null hypothesis leads to a small improvement in performance. However, in absolute terms we report uniformly low classification rates in predicting cluster membership as a proxy for expression profile or TF binding on all datasets. For synthetic data with know motifs, performance is far below the theoretical upper bounds given real-world sample sizes and sequence lengths.

Our results suggest discovering and validating motifs by attempting to predict expression patterns or TF binding computationally and without prior knowledge might be impossible due to the combination of several factors: the enormous size of the search space, the complexity of regulation, and the severely limited number of samples, namely the number of co-regulated genes identified with each expression profile cluster or the number of sequences bound by a given transcription factor. However, it must be emphasized this represents a *computational limitation*; since the theoretical upper bound is relatively high with real parameter settings, individual motifs could indeed be highly predictive were other biological signals incorporated into the discovery process.

Motif Discovery Strategies

Except for recent work, the previous literature on motif discovery is well-reviewed [8–10]. In principle, motif discovery need not be treated as a classification problem. Instead, generative sequence models can be constructed to represent DNA not involved in regulation, and one can then search for patterns in real sequences that are over-represented relative to this model with respect to some expression or transcription factor binding property. This is the approach of several methods [11–14] which seek the most significant ungapped local alignments over a range of lengths within a set of sequences. A common generative model for random DNA is a Markov chain. Let $X = \{X_1, X_2, ..., X_L\}$ be a DNA sequence

of length L, with $X_i \in \{A, C, G, T\}$. The m'th order Markov assumption is $P(X_i|X_{i-1}, X_{i-2}, ..., X_1) = P(X_i|X_{i-1}, X_{i-2}, ..., X_{i-m})$. However, Markov models fail to account for large-scale variation in low-order sequence properties such as AT content, making hypothesis testing based on such models of dubious value (see Results). They also fail to account for repetitive and transposable elements, which constitute a large fraction of certain genomes.

Once a list of candidate sequences, say of length W, is determined by some motif discovery algorithm (for example, aligned portions of each sequence in a local alignment-based algorithm), the standard way of generalizing to a sequence distribution P on W-mers is a position-weight matrix or PWM. The matrix is $4 \times W$: each row is a DNA nucleotide, each column a position, and the entry is the observed count in the set of sequences, possibly augmented with a pseudocount. Therefore each column represents the probability distribution of nucleotides at one of the W positions. Since the positions are assumed mutually independent, the log-likelihood of a W-mer is computed by summing the log-probabilities of the nucleotides observed at each position, which can be compared with the log-likelihood under some background model. If the background model also assumes mutually independent components, the likelihood ratio test is the Naive Bayes classifier for a fixed W-mer. When classifying an entire genomic sequence (e.g., the upstream region of a gene) according to membership in a gene cluster, high-likelihood W-mers under the PWM serve as features.

Another approach [15–18] is to search for motifs defined by an exact match to a short sequence or regular expression. This may be done discriminatively [15], i.e. treating motif discovery explicitly as a classification problem, or using a background model in a fashion similar to the alignment method described earlier. In either case statistical hypothesis testing may be performed [15, 17, 18] since the relative simplicity of the model makes this tractable. However, in the discriminative approach to discovery, there is generally no systematic validation on data not used to train the classifier, e.g., no mention of estimating performance with holdout data or cross-validation.

Some attempts [19–21] classify sequences based on kernels. In particular, the *spectrum kernel* of depth W of a sequence S records the number of occurrences of each of the 4^W possible W-mers in S; for instance, in [20], all 1024 DNA 5-mers are considered as features. These methods are well-grounded in statistical learning and properly validated. However, it is difficult to perform hypothesis testing or to interpret the decision-making in biological terms, such as TF binding, due to the number and diversity of features.

A few important contributions fall outside these broad categories. For example, DME [22] formulates a discriminative model by enumerative perturbation of a PWM; ANN-Spec [23] learns a neural network; and DEME [24] learns a discriminative PWM model with a conjugate gradient algorithm. None of these approaches addresses statistical hypothesis testing or error estimation by holdout testing or cross-validation. Seeder [25] is a method that minimizes a suitable distance between a seed and a larger sequence and bears some resemblance to the LR algorithm (see Methods) introduced here. Whereas hypothesis testing and false discovery rates are explicitly addressed, there is no consideration of holdout-validation or cross-validation.

Methods

Datasets and Preprocessing

We consider three datasets: Yeast expression profile clusters from Beer et al. [4]; human gene expression data from the Connectivity Map project [26]; and ChIP-chip transcription factor binding data from Harbison et al. [27]. Here, an expression profile is the collection of mRNA concentrations for a set of transcripts (or a surrogate for this quantity, such as microarray hybridization intensity) under a predetermined set of conditions and a cluster of genes corresponds to a coarse quantization of profiles. The upstream regions of a cluster of coexpressed genes are assumed to be enriched for active binding

sites (those that are accessible and actually bound *in vivo* under the relevant cellular conditions) for the transcription factor(s) responsible for the coexpression. Inactive TFBS, e.g. those not accessible to transcription factors due to chromatin structure or those that occur in the wrong context to modulate expression, are assumed to occur no more frequently in the upstream sequences of coexpressed genes than in randomly chosen upstream sequences. For the Beer data, 49 expression profile-based clusters were already specified based on the K-means algorithm [28]; we searched for motifs in the first 800 nucleotides upstream of the coding start site of each gene. For the Connectivity Map data, we generated 100 expression profile clusters using the K-means algorithm with Kendall's Tau [29] as a distance metric, and examined the first 2000 nucleotides upstream of the most upstream coding start site of each gene for motifs. For the Harbison et al. data, only the rich media data (binding affinities of 175 TFs) were used and the set of sequences binding to a given TF was treated as a cluster. Genome annotations were obtained from the UCSC Genome Browser [30]. In all cases clusters of fewer than ten genes are excluded from the analysis due to excessively small sample size. Table 1 summarizes the key properties of each dataset.

Classification Benchmark

Motif discovery algorithms can be validated by first grouping genes into disjoint clusters based on either similar expression profiles or common transcription factor binding and then attempting to assign cluster membership based on the existence of motifs. To quantify the regulatory predictive value of discovered motifs, for each cluster k we estimated the accuracy of the motif learned using k as the foreground cluster in discriminating between members and non-members of cluster k. This was done using holdout validation, or validation on labeled data disjoint from the training data. Each cluster was partitioned into two disjoint and equally-sized subsets, one for training and one for validation. When discovering a motif in cluster k, 200 sequences (100 training, 100 validation) randomly selected from the union of all clusters except k served as the "background" cluster. For the Harbison et al. data, since the clusters were not perfectly disjoint, sequences that appeared in the foreground cluster were excluded from the background cluster. For foreground clusters larger than 200 sequences, we randomly sampled 200 sequences to represent the cluster, for computational reasons. We used area under ROC curve (AUROC), a common performance metric in statistical learning, to measure the level of discrimination of the motif-based classifier for each cluster k.

All methods tested return a set of sequences of fixed length W that represent samples from the motif discovered. (The ALR method also returns information about differences in the bulk distribution between foreground and background.) For all methods a PWM with a pseudocount of 1 was built from the set of sequences returned. This was used to produce the PWM score for each sequence x. For simplicity, the PWM score was defined simply as the maximum likelihood of any subsequence of x or its reverse complement, $RC(x)$. The PWM score F can be described as the maximum of the following two quantities:

$$\max_{l=1,\ldots,N-W+1} \prod_{i=l,\ldots,l+w-1} PWM(x_i, i-l+1), \quad \max_{l=1,\ldots,N-W+1} \prod_{i=l,\ldots,l+w-1} PWM(RC(x)_i, i-l+1)$$

For all methods except ALR, the ROC curve and the AUROC value are generated by thresholding F. For the ALR method, F is combined with the bulk sequence features in a logistic regression model and the probability predicted by this model is thresholded to form the ROC curve. (The details of the ALR model are included later in this section.)

Planted Motif Simulations

The objective is to measure the performance of *de novo* methods, as well as the theoretical upper bound, when the true motif is generated from a realistic PWM and the complexities of the background model

are removed. We used *Saccharomyces cerevisiae* PWMs obtained from JASPAR CORE [31] with width W of at least eight nucleotides. These motifs were randomly planted in background DNA with sequence lengths and cluster sizes identical to the Beer et al. data, but generated from a zero-order Markov (or independent nucleotide) model with GC fraction equal to 0.4. For each cluster, a different motif was randomly chosen from among the available JASPAR motifs, and then inserted at a randomly chosen position in each member sequence by sampling independently from the distribution represented by each column of the PWM.

The theoretical upper bound on the AUROC performance of *de novo* methods was computed by assuming the true PWM was discovered and applying the classification benchmark described above. We used the same sets of foreground and background sequences for this benchmark as for the *de novo* benchmarks and reported the results on the sequences used for holdout in the *de novo* benchmarks so that the results would be directly comparable.

We also computed the accuracy of the Bayes rule for the full multiclass problem, i.e., for predicting which of the 49 clusters each sequence belonged to. We did this under the assumption of equal prior probabilities for all clusters. Since the cluster sizes vary significantly, using the true proportions would be overly optimistic. Since the positions of the motif were generated independently and every nucleotide of the background DNA was generated independently, the conditional independence assumption made by the naive Bayes classifier is in force, and hence the accuracy of naive Bayes is in fact the best possible (Bayes optimal).

Specifically, assume a sequence x of length N is scored for membership in cluster k. If x is a member of cluster k then a motif of length m represented by PWM_k has been planted starting at a random position L in x. The rest of x was generated independently with each nucleotide having probability $P_0(.)$. Let $Y \in \{1, ..., 49\}$ denote the true cluster. The probability that x is a member of k is proportional to the sum over all positions in x of the probability of observing x given that the motif represented by PWM_k was planted at positions $R_L = \{L, L+1, ..., L+m-1\}$:

$$\begin{aligned} P(Y = k|x) &\propto \sum_{j=1,...,N-m} P(x|Y = k, L = j) \\ &= \sum_{j=1,...,N-m} \prod_{i \in R_j} PWM_k(x_i, i-j+1)/N_{PWM_k} \prod_{i \notin R_j} P_0(x_i) \end{aligned}$$

Here, N_{PWM_k} is a normalizing constant representing the number of sequences PWM_k was built from, plus the pseudocount. This can be computed exactly for any cluster k and sequence x and the Bayes classifier is simply $f_B(x) = \arg\max_k P(Y = k|x)$, whose accuracy on the simulated dataset can also be easily computed.

Logistic Regression

We introduce a new method for motif discovery based on logistic regression (the LR algorithm), which allows rigorous hypothesis testing, including multiple testing correction, but only utilizes real sequences, not generative background models.

Given two clusters i, j, the goal is to find a PWM motif of some pre-specified width W that discriminates members of $\mathcal{S}_i = \{S_{i,1}, S_{i,2}, ..., S_{i,N_i}\}$ (the foreground cluster) from members of $\mathcal{S}_j = \{S_{j,1}, S_{j,2}, ..., S_{j,N_j}\}$ (the background cluster). (This method can be generalized to the case where the exact width of the motif is unknown by running it for multiple values of W and choosing the most significant result.) Let s be some fixed sequence of length W representing a candidate "core" motif and let $RC(s)$ denote its reverse complement. The distance $H(u, s)$ between s and any other sequence u of length W, is taken to be the minimum of the Hamming distances from u to s and from u to $RC(s)$. For any larger sequence S containing s, define

$$m(S, s) = min_{u \subset S, |u|=W} H(u, s) \quad (1)$$

which is minimum distance to s within S. Consider the set of distances to the sequences in $\mathcal{S}_i$:

$$m_{\mathcal{S}_i,s} = \{m(S_{i,1}, s), m(S_{i,2}, s), ..., m(S_{i,N_i}, s)\} \quad (2)$$

and similarly for $m_{S_j,s}$. We refer to these as "mismatch" statistics. These are the features we regress upon.

Now consider a logistic regression model designed to discriminate between members of $\mathcal{S}_i$ and $\mathcal{S}_j$ based on the mismatch features $m_{\mathcal{S}_i,s} \cup m_{\mathcal{S}_j,s}$. Let $Z = 1$ for sequences in $\mathcal{S}_i$ and $Z = 0$ for sequences in $\mathcal{S}_j$. For each fixed sequence (candidate core motif) s, we regress $\log \frac{P(Z=1|s)}{P(Z=0|s)}$ on the mismatch for s. The parameters are learned by maximum likelihood. Let M_s denote the model, which includes an intercept coefficient α and a mismatch coefficient β_s, and let L_s denote the log likelihood of the data under the model. We also learn a null model M_0 that includes only the intercept coefficient; let the likelihood of the data under this model be L_0..

Since M_0 is nested in M_s (the two models become equivalent if $\beta_s = 0$), we use Wilks' Theorem [32] to test M_0 vs M_s for each s of width W. The test statistic is $T_s = 2(L_s - L_0)$ which is asymptotically χ^2_1-distributed under M_0 since constraining $\beta_s = 0$ removes one degree of freedom. This procedure can be repeated either for every length W sequence or (for computational reasons) some subset thereof. In this paper every length W subsequence in $\mathcal{S}_i$ is used and the element of $\mathcal{S}_i$ that the current s was obtained from is excluded from the hypothesis testing stage to avoid bias. This procedure results in a well-defined number of hypothesis tests, each assigned a p-value p_s. (In fact, a false discovery rate [33] can be computed for any set of discovered motifs.) We only keep the most significant motif W-mer s^*, and learn a PWM from all length W subsequences s in any member of $\mathcal{S}_i$ such that $H(s^*, s) < median(m_{\mathcal{S}_j,s^*})$. In other words, the idea is to create a PWM from all foreground subsequences that match s^* better than any subsequence in the majority of sequences in the background set as assessed by median Hamming distance to s^*.

Computing $m(S, s)$ can be accelerated by preprocessing each S into a trie (also known as a prefix tree). The trie can be built in $O(|S||s|)$ time but needs to be built only once for each S with the cost being amortized over multiple values of s. If $m(S, s) = H$, the worst-case time complexity of computing $m(S, s)$ after the trie is built, for a four-letter DNA alphabet, is $O(min(|S||s|, |s|^{H+1}4^H))$. In practice this is extremely efficient because H is usually small, and for large H the worst case time complexity is equivalent to the naive algorithm of directly computing the Hamming distance between s and every length s subsequence of S.

The above model can be generalized to account for systematic variation in low-order sequence properties across clusters. We refer to this as the adjusted LR or ALR method since we attempt to find the most significant s *conditioned* on these low-order properties. Start with a set of some pre-determined size N_{spec} of features chosen from the feature space of low-order ($W = 1$ and $W = 2$) spectrum kernels. The set of features used is the N_{spec} features that are individually most predictive, with a sequence and its reverse complement treated as identical. More precisely, define X_{all} as the feature space of the union of spectrum kernels of $W = 1$ and $W = 2$, and let X_i represent the ith feature. For each i regress Z (as defined above) on only X_i and an intercept term and choose the N_{spec} features that produce the largest likelihoods in such models. Biologically, these are intended to represent bulk properties of the relevant DNA sequence, such as nucleosome affinity, melting ability and flexibility. Such low-order properties are individually significantly correlated with cluster membership. (See Results)

Finally, we redefine the null model M_0 to include both the intercept coefficient and β coefficients for N_{spec} low-order spectrum kernel features. The model M_s contains all of these features and additionally β_s for some length W sequence s. *The idea is to test whether anything additional is learned by adding information about s after accounting for low-order sequence properties.* M_0 is still nested in M_s and Wilks' Theorem can be used in the same way as above. A PWM is built from the highest scoring s as described above, and a final decision rule is learned that combines the spectrum kernel features with the PWM score. This is done by creating a logistic regression model with an intercept coeffient, one β

coefficient for each of the N_{spec} low-order features and one β coefficient for the PWM score of the PWM built from s^*.

A/T Fraction Test

We performed a *Monte Carlo* test to assess whether Markov models are "too null" by testing whether A plus T fraction varies more across different contiguous genomic sequences than could be explained by sampling variance under a single global Markov model of non-coding DNA. Let $\mathcal{U}_{real} = \{S_{1,real}, S_{2,real}, ..., S_{n,real}\}$ be a set of real DNA sequences used to train an m'th order Markov model, assumed to accurately capture the low-order properties of all sequences in $\mathcal{U}_{real}$.

Now, define $\mathcal{U}_j$ for $j \in \{1, 2, ..., K\}$ to be a set of synthetic sequences $\{S_{1,j}, S_{2,j}, ..., S_{n,j}\}$ sampled from this Markov model, where $|S_{i,j}| = |S_{i,real}|$. Define $AT(.)$ as the fraction of A plus T nucleotides in a given DNA sequence. For a sequence sampled from a Markov model with biologically realistic parameters, the A/T fraction will be approximately normally distributed. Let μ_i and σ_i be the mean and standard deviation, respectively, of $\{AT(S_{i,1}), AT(S_{i,2}), ..., AT(S_{i,K})\}$. If a single Markov model provides an adequate description of the low-order properties of real non-coding sequences, $Z_i = \dfrac{S_{i,real} - \mu_i}{\sigma_i}$ should be distributed approximately $N(0, 1)$.

Results

Even High-Order Markov Generative Models are Too Null

Monte Carlo simulations were performed to quantify the extent to which high-order Markov background models capture the structure of randomly selected sets of DNA sequences. MEME [11] was run approximately 15,000 times. We set the parameters to search for motifs of width $6 \leq W \leq 12$ and to not search for multiple occurrences of a motif, but rather to only consider two possibilities: the existence of one motif or none. MEME was applied to random gene sets from the union of the Beer et al. [4] clusters. The size of each gene set was chosen as half the size of a randomly chosen cluster from [4]. (The other half of the data was used to verify that the motif discovered does not predict cluster membership on held-out data, or in other words that our "null" model is actually null.) This protocol allows for a null model where random sets of input sequences were used, but the individual sequences and the cluster sizes were real. For this analysis, the sequences are the upstream region of each gene from the coding start site to the next upstream transcript on the same chromosome.

The background model used was the 6'th order Markov model of yeast intergenic regions, which is included with MEME. MEME reports an E-value, which is the expected number of motifs at least as strong as the one observed under the null model, and which always exceeds the corresponding p-value. In this case the null (respectively, alternative) hypothesis is that no (resp., at least one) over-represented motif exists in the gene set. Under the null, p-values should be uniformly distributed over $[0, 1]$. A similar test was performed using the LR algorithm with the same data set, as well as the $2,000$ nucleotides upstream of the coding start site for genes in the Human Cmap dataset. Disjoint random gene sets were used for $\mathcal{S}_i$ and $\mathcal{S}_j$. The FDR as computed by the LR algorithm also controls the family-wise error rate (the probability of making at least one Type I error) if all null hypotheses are true [33], as is the case in this *Monte Carlo* simulation. Therefore, an estimated false discovery rate Q^* for the most significant motif should occur with probability no greater than Q^*. Under our null model of random sequence clusters, the E-values reported by MEME are anti-conservative even if interpreted as p-values (Figure 1a). Figures 1b and 1c demonstrate that false discovery rates produced by the LR algorithm are approximately accurate when all null hypotheses are true.

As described in Methods, a test for overdispersion of A/T fraction relative to the variance expected under a high-order Markov model was also performed. We used the first $2,000$ nucleotides upstream of the

transcription start site for each gene in the human Connectivity Map [26] gene set and the full upstream intergenic regions of all yeast genes in the Beer et al. [4] dataset, ignoring regions with fewer than 100 nucleotides. A 6'th order Markov model was applied to all yeast sequences and another one for all human sequences. If all upstream sequences in each dataset were well-approximated by a single Markov model, the distribution of Z-scores would be expected to be approximately $Normal(0, 1)$. However, this is not the case (Figure 1d). There are at least two possible explanations: either even a 6'th order Markov model does not capture low-order properties of intergenic sequences, or such a model would not have homogeneous parameters across genomic regions.

Finally, low-order sequence properties differ significantly across clusters, as illustrated in Table 2, further supporting the hypothesis that commonly used generative background models fail to capture important structure. The fraction of each dimer in each sequence was regressed on cluster membership, with each dimer and its reverse complement being treated as identical. The value of R^2 is the fraction of total variance in the frequency distribution of each dimer that is explained by cluster membership. The p-value tests the null hypothesis that the mean frequency of the dimer in each sequence is identical across clusters.

The overall conclusion of these experiments is that common generative approaches may produce overly optimistic conclusions about the significance of motifs discovered because the null models are "too null", or assume that bulk non-coding DNA is more random than it really is.

Classification Rates of all Tested Methods are Poor

We evaluated the performance of five motif discovery algorithms in predicting gene cluster membership on three real datasets and one synthetic dataset. The methods are MEME [11], $6 \leq W \leq 12$, zero or one motif instance per sequence model; AlignAce [12], numcols = 12; DEME [24], $W = 12$; and the LR and ALR algorithms, with $6 \leq W \leq 12$ and $N_{spec} = 3$ for ALR. Recall that the ALR algorithm accounts for the discriminating power of bulk DNA features, whereas the LR algorithm ignores low-order features and hence can be compared directly with the other three methods. The datasets are Human Cmap, yeast Beer et al., and yeast Harbison et al., and the simulated yeast dataset mentioned previously. In all cases both the forward and reverse complement strand were analyzed. For MEME, the background model was second-order Markov and estimated from the background sequences. For AlignAce, the background GC content was set to that the background sequences. Whenever a method output a list of sequences, these were converted to a PWM by calculating the empirical distribution over the four nucleotides at each position, except a unit pseudocount was incorporated to avoid zero probabilities. Table 3 displays the estimated mean holdout AUROC values across clusters for all datasets and algorithms. Table 4 displays the resubstitution AUROC.

There is a wide range of classification rates from method to method and from dataset to dataset. The best performance in all real datasets is obtained by the ALR algorithm. This may be due to the regression framework, which accommodates hypothesis testing without a synthetic background model and learns bulk sequence properties in addition to PWMs. DEME, which also takes an explicitly discriminative approach to motif discovery, also fares relatively well compared with AlignAce and MEME, which use artificial background null models. However, the most striking trend is that all mean AUROCs are disappointing, reaching at most around 0.62 on real data and 0.72 on synthetic data.

To determine the theoretical upper bound classification accuracy we computed the mean AUROC on the synthetic dataset as in Table 3 but used the PWM motifs that we planted instead of discovered PWMs as classifiers. The mean AUROC here was 0.865. Furthermore, the multi-class Bayes accuracy for determining the correct cluster membership for each sequence from all 49 available clusters was 0.344. These results demonstrate the substantial discriminative ability of single motifs in this context and the theoretical possibility of achieving much higher one-versus-all discriminative power than any method achieved on any dataset.

Proper Hypothesis Testing Predicts Generalization

The purpose of hypothesis testing is to determine which motifs may be biologically relevant signals and which are more likely statistical noise. We compared the mean AUROC for significant (FDR ≤ 0.05) vs. non-significant (FDR > 0.05) motifs using our LR and ALR methods and all datasets. Table 5 shows that statistical significance does predict generalization of a motif's discriminative ability from the training set to the validation set. Note that ALR retains substantial discriminative ability in the absence of a significant motif because the bulk features also contribute to discrimination. LR, on the other hand, performs barely better than the 0.5 AUROC expected under random guessing.

Discussion

This study has identified the following key findings: First, we have shown that even high-order generative models of random DNA are "too null" resulting in overly optimistic estimates of motif discovery in DNA sequences. Motif discovery methods should therefore be evaluated in a classification framework using only real DNA sequences. Second, we have shown that rigorous hypothesis testing can still be incorporated by providing a well-defined null hypothesis, namely that a motif is over-represented in a well-defined foreground cluster relative to a well-defined background cluster, where again both consist of real sequences. Given that this approach is discriminative, standard methods of validating a classifier (namely holdout or cross-validation) can and should be used. Applying this stringent benchmark, all methods perform poorly regardless of whether sequences are clustered by ChIP-chip TF binding or by mRNA expression. Explicitly discriminative methods and those that account for differences in bulk DNA properties have marginally improved discriminative ability. Thus, there is clearly much need for improvement in computational discovery of motifs. Given the statistical difficulties, it is clear that biological knowledge and the integration of high-throughput and heterogeneous data will not only be useful, but essential to ultimately achieving high accuracy. For example, accounting for chromatin modification [34] might aid in distinguishing between TFBS that are accessible and therefore potentially bound and those that are inaccessible to the relevant TFs under a given cellular condition.

The relatively high theoretical upper performance bound on the synthetic data suggests that the poor performance of *de novo* methods cannot be attributed primarily to low predictive value of individual motifs. Thus, the relatively poor performance on real data of methods that focus on finding individual motifs may not be, of itself, strong evidence that individual motifs are not biologically prevalent and meaningful in vivo. The gap between this upper bound and the performance of *de novo* discovery algorithms on the synthetic dataset is also much larger than the performance gap between *de novo* discovery on the real Beer et al. vs. synthetic datasets. The synthetic dataset has sequence lengths and cluster sizes identical to the Beer et al. data but conforms to our simplified biological model that exactly one motif exists in each cluster and that the motif is reasonably long and can be statistically modeled accurately by a PWM. These results suggest that statistical tractability is a more severe problem than deficiencies in the biological models used by discovery algorithms, such as ignoring combinatorial regulation or using PWMs instead of more complex models of motifs. Using more complex, biologically realistic models such as attempting to simultaneously discover motifs involved in combinatorial regulation or removing the assumption of conditional independence between positions that the PWM model implies would likely exacerbate these statistical issues. This is especially true since the sample size available for discovery is limited by the number of occurrences of a given cis-regulatory element in the genome. This limit can be effectively increased by using phylogenetic methods such as PhyloGibbs [35]. However, this assumes that cis-regulatory elements are mostly conserved across species and only increases sample size incrementally.

It might be argued that clusters based on expression of downstream genes or ChIP-chip TF binding are noisy representations of the set of sequences regulated by a given cis-regulatory mechanism, e.g. the

binding of a specific TF. For example, TF binding has been shown to be context specific with regard to developmental stage [36] and ChIP data does not always accurately predict transcription factor binding events catalogued in the literature [37]. However, poor performance is observed regardless of whether clusters are defined by ChIP-chip transcription factor binding or expression clustering. Furthermore, results on the synthetic data, which uses planted motifs and thus guarantees that every member of a given cluster contains an example of the same TFBS, are only incrementally stronger than those on the Beer et al. data. Taken together, our results suggest that inaccuracies in clustering caused by imperfections in expression or ChIP-chip data cannot fully explain the poor performance. Furthermore, we argue that imperfect clusters represent a more realistic use case for *de novo* discovery methods than the ideal case where each cluster perfectly represents the set of sequences regulated by a given cis-regulatory mechanism, making our discriminative benchmark highly relevant.

The strong resubstitution performance (i.e. when the model is trained and tested on the same data), especially on the synthetic data, suggests that statistical tractability is also a more severe problem than optimization of the objective functions of the discovery algorithms. Such discrepancies between resubstitution error and error in a test set are clear indicators of overfitting issues.We have established an upper performance bound of approximately 0.865 on the synthetic dataset. All algorithms achieve an average resubstitution AUROC of at least 0.8 on this dataset, even though all use heuristics to optimize their objective functions. In other words, all algorithms on average find solutions with discriminative power comparable to the correct solution on the training data even though these often don't generalize – as seen by greatly reduced AUROC in holdout validation.

The difficulty of developing adequate generative models of background DNA sequences may be partly due to the variation in low-order sequence properties across expression and transcription factor binding profiles. This phenomenon has previously been observed specifically with respect to GC content in human promoter regions [38]. Our results suggest that the phenomenon of low-order, large scale sequence properties being correlated with expression and TF binding is more broadly applicable, both to yeast and to the frequencies of a variety of dimers. They also suggest that the statistical significance of longer (e.g. 12-mer) motifs may sometimes result from of variation in the low-order properties of sequences across expression or TF binding profiles, and these would not be significant *given* the low order properties of the sequences in which they were found. Such motifs are not likely to be biologically meaningful.

Nucleosome positioning can be predicted by low-order sequence properties such as 1- through 4-mer frequences [1]. Nucleosomes appear to be depleted in active regulatory regions [39]. This observation and a study in the PH05 promoter [40] suggest that high nucleosome occupancy may interfere with transcription factor binding in at least some cases. If the main link between low-order sequence properties and gene expression were that low-order sequence properties affect nucleosome occupancy, which affects TF binding and in turn affects expression, then low-order sequence properties in small, specific regions would be expected to predict expression and TF binding, but not low-order properties across large (several hundred nucleotides or more) regions. Similarly, such properties would not be expected to be frequently predictive in regions such as introns, where TFBS occur less frequently than near the transcription start site (TSS).

Conclusion

De novo DNA motif discovery remains a challenging problem with computational methods. The fact that common generative models, or explicit models of the probability distribution of "random" DNA, cannot represent "random" DNA with reasonable accuracy motivates a discriminative approach where real background sequences are used. However, applying traditional validation protocols for classification algorithms reveals universally disappointing rates in predicting expression profile or TF binding from sequence. The largest obstacle may be over-fitting, which will be difficult to overcome since the samples size is effectively the number of genes in strongly co-regulated clusters or bound by a given TF, and thus

cannot be expanded arbitrarily to provide the necessary statistical power.

Acknowledgments

References

1. Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, et al. (2007) Nucleosome positioning signals in genomic DNA. Genome Research 17: 1170 –1177.
2. Mandel M, Marmur J, Grossman KML (1968) Use of ultraviolet absorbance-temperature profile for determining the guanine plus cytosine content of DNA. In: Nucleic Acids Part B, Academic Press, volume Volume 12, Part 2. pp. 195–206.
3. Straussman R, Nejman D, Roberts D, Steinfeld I, Blum B, et al. (2009) Developmental programming of CpG island methylation profiles in the human genome. Nat Struct Mol Biol 16: 564–571.
4. Beer MA, Tavazoie S (2004) Predicting gene expression from sequence. Cell 117: 185–198.
5. Salem A, Ray DA, Xing J, Callinan PA, Myers JS, et al. (2003) Alu elements and hominid phylogenetics. PNAS 100: 12787–12791.
6. Kamal M, Xie X, Lander ES (2006) A large family of ancient repeat elements in the human genome is under strong selection. PNAS 103: 2740–2745.
7. Narlikar L, Gordân R, Hartemink AJ (2007) Nucleosome occupancy information improves de novo motif discovery. In: Proceedings of the 11th annual international conference on Research in computational molecular biology. Berlin, Heidelberg: Springer-Verlag, RECOMB'07, pp. 107–121.
8. Tompa M, Li N, Bailey TL, Church GM, Moor BD, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotech 23: 137–144.
9. Sandve G, Abul O, Walseng V, Drablos F (2007) Improved benchmarks for computational motif discovery. BMC Bioinformatics 8: 193.
10. Das M, Dai H (2007) A survey of DNA motif finding algorithms. BMC Bioinformatics 8: S21.
11. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucl Acids Res 34: W369–373.
12. Roth FP, Hughes JD, Estep PW, Church GM (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nature Biotechnology 16: 939–945.
13. Liu X, Brutlag DL, Liu JS (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput 2001 : 127–138.
14. Frith MC, Saunders NFW, Kobe B, Bailey TL (2008) Discovering sequence motifs with arbitrary insertions and deletions. PLoS Comput Biol 4: e1000071.
15. Sinha S (2003) Discriminative motifs. Journal of Computational Biology 10: 599–615.
16. Pavesi G, Mereghetti P, Mauri G, Pesole G (2004) Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucl Acids Res 32: W199–203.

17. Sinha S, Tompa M (2003) YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucl Acids Res 31: 3586–3588.

18. Marschall T, Rahmann S (2009) Efficient exact motif discovery. Bioinformatics 25: i356 –i364.

19. Leslie C, Eskin E, Noble W (2002) The spectrum kernel: A string kernel for SVM protein classification. Pac Symp Biocomput 2002 .

20. Vert JP, Thurman R, Noble WS (2005) Kernels for gene regulatory regions. In: NIPS'05.

21. Lee D, Karchin R, Beer MA (2011) Discriminative prediction of mammalian enhancers from DNA sequence. Genome Research 21: 2167–2180.

22. Smith AD, Sumazin P, Zhang MQ (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. PNAS 102: 1560 –1565.

23. Workman CT, Stormo GD (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. Pac Symp Biocomput 2000 : 467–478.

24. Redhead E, Bailey TL (2007) Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. BMC bioinformatics 8:385.

25. Fauteux F, Blanchette M, Strmvik MV (2008) Seeder: discriminative seeding DNA motif discovery. Bioinformatics 24: 2303 –2307.

26. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, et al. (2006) The connectivity map: Using Gene-Expression signatures to connect small molecules, genes, and disease. Science 313: 1929 –1935.

27. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99–104.

28. MacQueen J (1967). Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281-297.

29. Kendall MG (1938) A new measure of rank correlation. Biometrika 30: pp. 81-93.

30. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, et al. (2003) The UCSC genome browser database. Nucleic Acids Research 31: 51 –54.

31. Bryne JC, Valen E, Tang ME, Marstrand T, Winther O, et al. (2007) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Research 36: D102–D106.

32. Wilks SS (1938) The Large-Sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics 9: 60–62.

33. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57: 289–300.

34. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, et al. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Research 21: 447–455.

35. Siddharthan R, Siggia ED, van Nimwegen E (2005) Phylogibbs: A gibbs sampling motif finder that incorporates phylogeny. PLoS Comput Biol 1: e67.

36. Wilczynski B, Furlong EEM (2010) Dynamic CRM occupancy reflects a temporal map of developmental progression. Molecular Systems Biology 6:383.

37. Chen K, van Nimwegen E, Rajewsky N, Siegal ML (2010) Correlating Gene Expression Variation with cis-Regulatory Polymorphism in Saccharomyces cerevisiae. Genome Biology and Evolution 2: 697–707.

38. Saxonov S, Berg P, Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. PNAS 103: 1412–1417.

39. Lee C, Shibata Y, Rao B, Strahl BD, Lieb JD (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat Genet 36: 900–905.

40. Svaren J, Hrz W (1997) Transcription factors vs nucleosomes: regulation of the PH05 promoter in yeast. Trends in Biochemical Sciences 22: 93–97.

Figure Legends

Figure 1. Generative models are too null. Panel (a): Quantile plot of Meme E-values for approximately 15,000 random runs, with E-values > 1 excluded. The X-axis represents the log_{10} E-value as reported by MEME. The Y-axis represents the log_{10} quantile. For example, under our null model E-values below 10^{-10} are reported with probability slightly more than 10^{-2}. Panels (b) and (c): Quantile plots of LR false discovery rates, similar to the Meme E-value quantile plots, for the Beer et al. and Human Cmap datasets respectively. Panel (d): Z-score plots of A/T fraction of yeast and human intergenic sequences relative to the distribution expected under a 6th order Markov model, with the standard normal distribution (red) shown for reference.

Tables

Dataset	N Clusters	N Seqs	Clustering Method	Sequences
Beer et al.	49	48	K-means, Pearson Correlation	800 BP upstream of coding start
Harbison et al.	175	128	ChIP-Chip TF Binding	Binding seqs provided by Harbison et al.
Human CMap	100	100	K-Means, Kendall's Tau	2000 BP upstream of coding start

Table 1. The properties and sizes of the datasets used. N Seqs is the number of clusters that contained at least ten sequences. Clusters with fewer than ten sequences were excluded from the analysis due to excessively small sample size.

Dimer	Beer et al.	Harbison et al.	Human Cmap (Upstream)	Human Cmap (Introns)
AA/TT	0.076 (8.94e-21)	0.083 (4.47e-77)	0.110 (8.39e-208)	0.267 (0)
AC/GT	0.053 (7.38e-11)	0.057 (1.27e-32)	0.030 (6.09e-29)	0.068 (2.02e-102)
AG/CT	0.033 (0.000485)	0.056 (1.94e-30)	0.076 (2.24e-127)	0.210 (0)
AT	0.070 (6.34e-18)	0.148 (1.33e-222)	0.088 (1.84e-155)	0.231 (0)
CA/TG	0.037 (3.12e-05)	0.056 (5.35e-30)	0.078 (5.43e-131)	0.141 (2.01e-277)
CC/GG	0.115 (2.73e-40)	0.156 (7.28e-245)	0.101 (1.28e-186)	0.228 (0)
CG	0.093 (4.85e-29)	0.158 (2.15e-249)	0.081 (1.85e-139)	0.095 (4.04e-166)
GA/TC	0.047 (1.44e-08)	0.051 (1.25e-22)	0.041 (2.92e-49)	0.164 (0)
GC	0.078 (1.36e-21)	0.149 (1.56e-227)	0.081 (3.88e-139)	0.207 (0)
TA	0.051 (5.05e-10)	0.131 (5.53e-183)	0.098 (1.37e-180)	0.265 (0)

Table 2. The fraction of variance in dimer frequency across sequences explained by expression profile or transcription factor binding sequence set and associated F statistic P-value. For the Human Cmap data, this was assessed both for the 2,000 nucleotides upstream of the coding start site and for the intron sequences.

	Mean AUROC (Holdout)			
	Beer et al.	Harbison et al.	Human CMap	Synthetic
LR	0.591	0.600	0.530	0.677
ALR	0.620	0.629	0.569	0.683
MEME	0.598	0.536	0.521	0.718
AlignAce	0.561	0.524	0.524	0.660
DEME	0.613	0.557	0.541	0.677

Table 3. The mean AUROC of all algorithms on all datasets using independent holdout data. This validation is unbiased.

	Mean AUROC (Resubstitution)			
	Beer et al.	Harbison et al.	Human CMap	Synthetic
LR	0.776	0.771	0.731	0.814
ALR	0.836	0.857	0.799	0.858
MEME	0.753	0.784	0.637	0.809
AlignAce	0.657	0.693	0.584	0.831
DEME	0.835	0.848	0.799	0.894

Table 4. The mean AUROC of all algorithms on all datasets based on training and testing on the same data. The optimistic bias reveals massive overfitting.

	Mean AUROC (Non-Significant/Significant)			
	Beer et al.	**Harbison et al.**	**Human CMap**	**Synthetic**
LR	0.531/0.722	0.562/0.656	0.510/0.569	0.536/0.796
ALR	0.571/0.727	0.580/0.697	0.562/0.587	0.521/0.790

Table 5. The mean holdout AUROC of the LR and ALR algorithms for motifs for non-significant (FDR > 0.05) and significant (FDR ≤ 0.05) motifs respectively.