DISCOVER: a feature-based discriminative method for motif search in complex genomes

Wenjie Fu‡, Pradipta Ray† and Eric P. Xing§
School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA

ABSTRACT

Motivation: Identifying transcription factor binding sites (TFBSs) encoding complex regulatory signals in metazoan genomes remains a challenging problem in computational genomics. Due to degeneracy of nucleotide content among binding site instances or motifs, and intricate ‘grammatical organization’ of motifs within cis-regulatory modules (CRMs), extant pattern matching-based in silico motif search methods often suffer from impractically high false positive rates, especially in the context of analyzing large genomic datasets, and noisy position weight matrices which characterize binding sites. Here, we try to address this problem by using a framework to maximally utilize the information content of the genomic DNA in the region of query, taking cues from values of various biologically meaningful genetic and epigenetic factors in the query region such as clade-specific evolutionary parameters, presence/absence of nearby coding regions, etc. We present a new method for TFBS prediction in metazoan genomes that utilizes both the CRM architecture of sequences and a variety of features of individual motifs. Our proposed approach is based on a discriminative probabilistic model known as conditional random fields that explicitly optimizes the predictive probability of motif presence in large sequences, based on the joint effect of all such features.

Results: This model overcomes weaknesses in earlier methods based on less effective statistical formalisms that are sensitive to spurious signals in the data. We evaluate our method on both simulated CRMs and real Drosophila sequences in comparison with a wide spectrum of existing models, and outperform the state of the art by 22% in F1 score.

Availability and Implementation: The code is publicly available at http://www.sailing.cs.cmu.edu/discover.html.

Contact: epxing@cs.cmu.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Deciphering the gene control circuitry encoded in the genome is a fundamental problem in developmental biology (Micheloson, 2002). In multi-cellular eukaryotic organisms such as the metazoans, the time- and tissue-specific expression of essential genes during various developmental and physiological processes is carried out by an intricate interplay between the transcriptional factors (TFs), and their regulatory mechanisms which control the binding of the factors to recognition sites, known as TF binding sites (TFBSs), or motifs, within the regions of the DNA sequence called gene regulatory regions (Davidson, 2001). Motifs often appear as recurring, degenerate short string patterns (noisy copies of each other) in the non-coding, regulatory regions of the genome. It has been shown that in higher eukaryotes, instances of TFBS of each TF usually occurs clustered in several small regions of the genome (usually 200–2000 bp) known as cis-regulatory modules (CRMs) near the coding region of the gene being regulated. Each CRM typically contains more than one type of TFBS for implementing the logic required to regulate the gene correctly throughout the lifetime of the organism (Davidson, 2001).

Due to the degeneracy of the nucleotide content among motif instances, pattern matching-based in silico motif search in higher eukaryotes remains a difficult problem, even when using formalisms such as the position weight matrix (PWM) (or nucleotide distributions at each position of the motif).

The ‘grammatical organization’ of motifs within CRMs that encode complex spatio-temporal regulatory information can further complicate motif search compared with similar tasks in simpler organisms such as yeast (Frith et al., 2002). Extant methods based on simple pattern matching scores often yield a large number of false positives (FPs) (Sandve and Drablos, 2006), especially when the sequence to be examined spans a long region (e.g. tens of thousands of basepairs) beyond the basal promoters, where possible enhancers and CRMs could be located.

In this article, we concern ourselves with searching for instances of motifs and CRMs in higher eukaryotic genome based on not only a given description of the motif sequence patterns, such as the PWMs, but also additional features that distinguish a putative motif from the background. Our proposed approach is based on a discriminative probabilistic model known as conditional random field (CRF) that explicitly optimizes the predictive probability of motif presence in a large background, rather than the joint probability of both motif and background sequence under a generative model, as in many of the current methods reviewed below, whose predictive power can be seriously compromised when the amount of background sequence significantly dominates that of the motifs. See Figure 1 for a schematic workflow.

Numerous efforts have been made to predict CRMs comprising of a cluster of TFBSs (Berman et al., 2002), or to use cluster-based analyses to assist TFBS prediction. Some methods directly count the number of matches of some minimal strength to given motif patterns within a certain window of DNA sequence (Donaldson et al., 2005; Rajewsky et al., 2002; Rebeiz et al., 2002; Sharan et al., 2003). From a modeling point of view, this family of algorithms assumes that motifs are uniformly and independently distributed within a fixed size window. Such methods are conceptually straightforward and often simple to implement and computationally efficient. In practice, setting the optimal window size can be difficult and optimal

© 2009 The Authors
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
et al. Fig. 1. A schematic view of the workflow.

parameters may not be robust on input data and may require careful analysis to calculate (Lin et al., 2008). Further, an i.i.d. distribution of motifs is now known to be an unrealistic assumption (Bulyk et al., 2002).

A second major class of methods adopt a generative formalism to model the occurrences of motifs and CRMs as the output of some hidden stochastic processes, such as a first-order hidden Markov model (HMM), which removes the necessity of modeling the window size. The hidden-state transition matrix within the HMM usually corresponds to a set of soft constraints on the expected CRM length and the inter-CRM distance in terms of geometric distributions. HMMs that capture motif distributions, as well as intra-CRM and inter-CRM backgrounds, have been used in several prediction algorithms, e.g. Cister (Frith et al., 2002), and Cluster-Buster (Frith et al., 2003). Further extensions have been made to include distinct motif-to-motif transition probabilities in programs such as Stubb (Sinha et al., 2006), Module Sampler (Thompson et al., 2004) and BayCis (Lin et al., 2008), which also employs generalized, hierarchical HMMs. These extended models often require a significant amount of training data. Moreover, logical rules have recently been applied in a model on yeast data (Noto and Craven, 2007) in order to try and capture regulatory logic models in the spirit of Davidson (Davidson, 2001). While the HMMs and HMM-like models are capable of describing the architecture and properties of CRMs to a certain degree, the expressive power of HMMs is insufficient in that they cannot support complex representations for motifs such as non-local, sequence–composition based or epigenetic features surrounding the motif. As a result, their performances on complex CRMs such as those of the Drosophila early developmental genes are still unsatisfactory.

Phylogenetic conservation has been historically one of the most commonly used features used besides binding specificity to detect TFBSs (Loots et al., 2002; Moses et al., 2004). However, these algorithms are restricted to very closely related organisms (no more than 50 million years to the most recent common ancestor (Ray et al., 2008)), because non-coding sequences are difficult to align across large evolutionary distances due to commonplace evolutionary forces like duplication and shuffling in the regulatory genome, hence making orthology prediction difficult (Davidson, 2001). Several comparative genomic methods have been applied to CRM and motif prediction (Ray et al., 2008; Siddharthan et al., 2004; Sinha and He, 2007; Sinha et al., 2004). In this article, we concern ourselves only with motif detection within a single species, but we try and use additional features which use phylogenetic data from other species to analyze the effect of multi-species data on motif discovery.

A key motif representation used in all the above methods to score possible motif occurrence in an input DNA sequence is the PWM (Staden, 1984), also known as position-specific scoring matrix (PSSM) (see Supplementary Material for details). Several motif detection algorithms work based on designing hard constraints on features associated with motifs, like distance to transcription start site (TSS) (Sinha et al., 2008). Recently, there has been a number of works in the literature that focus on refining predictive models for individual TFBS by using a wide range of features that have been shown to correlate well with regulatory regions in general and with TFBSs in particular, without necessarily modeling the CRM structure (Nalikat et al., 2007; Naughton et al., 2006; Pudimat et al., 2004; Sharon and Segal, 2007). Using biologically motivated features like presence or absence of CpG islands, nucleosome sites, and helical structures, they appear to be able to significantly outperform models based on PWM motif representation alone. Pudimat et al. (2004) models a variety of features to assist in predicting binding sites, but selects the set of features in a greedy fashion, and models the features as nodes of a generic graphical model, causing topology selection of the graphical model to be NP-hard (Pearl, 1988). Sharon and Segal (2007) uses Markov networks to associate specific features with subsets of TFBS positions, causing the difficult problem of estimating the network structure to arise. Ernst (2008) analyzes a set of features to derive informative priors for TFBS prediction, using logistic regression-based classifiers for the choice of each feature. Such discriminative, integrative models have also achieved some success on other problems like protein fold recognition (Damoulas and Girolami, 2008).

In this article, we present DISCOVER : DIstributive COnditional random field for motif recoVERy in metazoan genomes. DISCOVER is a discriminative method for motif detection in higher eukaryotic genomes that enjoys the dual advantage of modeling CRM architecture of sequences and features of individual motifs. It is a CRF model (Lafferty et al., 2001), which incorporates a wide range of both CRM structure-based and individual motif-based features. CRF's have previously been used in sequence analysis, most notably in gene prediction (DeCapprio et al., 2007; Gros et al., 2007), since coding regions are much better characterized in terms of sequence level features with respect to regulatory regions. Bockhurst and Craven (2005) has applied a similar scheme to identify regulatory signals in prokaryotic sequences; but their model employs a simple feature set to resolve the motif sequence overlap problem, and also requires a pre-screening of motif scores via basic PWM-based models.

Our method is important in several respects in the context of the literature. First, it is a discriminative model explicitly tailored towards maximizing the conditional likelihood of predicting motifs, rather than maximizing the joint likelihood—which often confounds the analysis in the case of generative models. Secondly, it employs a comprehensive set of features carefully selected from the literature designed to capture a variety of characteristics of the motif and CRM patterns. Thirdly, it is an integrative model that allows sequence-specific features to be added at will to enhance the prediction scheme. Further, since feature scores are computed offline, it is easier
to incorporate scores involving complicated computation and long computation times as well as long-range dependencies.

We evaluate the CRF model on both simulated CRMs and actual biologically validated transcription regulatory sequences of *Drosophila melanogaster*, in comparison with a wide spectrum of existing models including, Cister (Frith et al., 2002), Cluster-Buster (Firth et al., 2003), BayCis (Lin et al., 2008), MSCAN (Alkema et al., 2004), Ahab (Rajewsky et al., 2002) and Stubb (Sinha et al., 2006). The results suggest that our proposed method significantly outperforms others on real *Drosophila* sequences.

The remainder of the article is outlined as follows: we discuss the model and feature design in Section 2. In Section 2.1, we describe how to learn the model from data, and then we briefly mention the inference algorithm given the model. Biological and empirical justifications for the features, experimental setup and results are presented in Section 3. We finish by some discussion on the scope of the model in Section 4.

2 METHODS

The conventional PWM representation for TFBSs is not discriminative enough to distinguish true binding sites from false binding sites. We desire a model for TFBSs and genomic sequence that supports a more complex motif representation without losing the ability to characterize sequence wide properties, which means a flexible feature design. The CRF model—a feature-based log-linear model in which features are easily incorporated—is an appropriate model choice under the circumstances. The basic inputs to such a computational model is a set of genetic sequences, a set of feature values corresponding to every nucleotide in the sequences and the PWMs of TFs that are being predicted. The output of the model is a prediction of a set of TFBSs which are being predicted, ranked in order of decreasing likelihood. The CRM boundaries can also be similarly predicted, but in this article we focus on the analysis of the TFBS predictions.

A CRF model that describes a conditional probability distribution of a genomic sequence is defined as:

\[
\begin{align*}
\mathbf{f}(y | x) &= \left[ \sum_{\lambda} f_{\lambda} (x | y) \right] \cdot S(\mathbf{x}, \mathbf{y}) \\
\text{where} \quad Z &= \left[ \sum_{\lambda} f_{\lambda} (x | y) \right] \cdot S(\mathbf{x}, \mathbf{y})
\end{align*}
\]

where we use \( x_i \) to represent the type of the observed nucleotide at site \( i \) in a sequence, and \( y_i \) to represent the hidden state associated with \( x_i \), which corresponds to the functionality of the site in the genomic sequence. The value of a hidden state is also called a state label. Vector \( x = [x_1, i = 1, 2, \ldots , L] \) and vector \( y = [y_1, i = 1, 2, \ldots , L] \), where \( L \) is the length of the sequence. Vector \( F \) is the set of features, each element \( F_i \) of which is the sum of feature scores of a particular feature category (where feature scores refer to the numerical value of the feature). Vector \( \mathbf{F} \) corresponds to the feature weights assigned to the set of features, and is learnt from data to decide which features may be more important in predicting TFBSs. \( Z \) is a partition function that normalizes the pdf and is a function of \( x \) and \( \lambda \). The value space for each \( x_i \) is \( \{A, C, G, T\} \). The values represent the four types of nucleotide in DNA, adenine, cytosine, guanine and thymine, respectively. The value space for hidden states \( y_i \) however, is not so straightforward, and it will be defined subsequently.

State design: we design a set of hidden states based on the possible functionality of each nucleotide in the genomic sequence being analyzed. We incorporate each motif type as a state since this is our prediction goal. We number the types of motifs and name the state for the \( m \)-th motif type \( M^{m} \).

Representatively, a hidden state \( y_i \) being state \( M^{m} \) implies that a motif of the \( m \)-th type is located starting at site \( i \) of the sequence. Those states are all that we need to represent binding sites. Next, we know that TFs are usually working together to regulate genes, especially in genomes of higher organisms. In order to work together, different types of TFBSs often lie close to each other in the range of hundreds of base pairs forming a so-called CRM (Davidson, 2001). We use state C to represent all nucleotides in the CRM regions except those binding sites which have already been labeled as Ms. The nucleotides which are still unlabeled after the first two rounds are set to state G, which represents a global background in the genomic sequence. Hence, the set of hidden states for modeling the functionality at a nucleotide position is given by \( S = \{G,C,M^{1},\ldots ,M^{n}\} \), where \( n \) is the number of motif types. We do not allow two motifs to share the same starting position, but such occurrences are infrequent. It is still an improvement on HMM-based approaches where modeling even partial overlap of motifs causes a combinatorial increase in the state space. Overlapping of starting positions of TFBSs can be accommodated in our model by using marginal probabilities in the prediction step.

Feature design: each element \( f(y_i | x_i) \) of vector \( \mathbf{f}(y | x) \) in Equation (1) is the sum of feature scores of a particular feature category, where feature score simply refers to the numerical value of the feature. It sums up feature function \( f_i \) over the sequence, which have a common meaning and share the same weight. An example is shown in Equation (16) of Supplementary Material, after we see some concrete features. The design of \( f_i \)’s is a critical part of CRF models. We include a rich set of features, most of which are introduced in Section 3. The set of features includes conventional features (TFBS sequence specificity, state transition probability) as well as evolutionary features (like presence of repeats, and of conservation across species), structural and epigenetic features (like melting temperature, nucleosome occupancy), features related to the protein coding mechanism (like distance to TSS, presence in 3′-UTR region), and additional discriminative features (like reverse complementarity of a site, and conservation symmetry). Their formal definitions can be found in Supplementary Material.

Features with a one-to-one correspondence with nucleotide base pairs can be easily integrated into the framework by defining as:

\[
\mathbf{f}(y_i | x_i) = \left( \sum_{m} f_{m}(x_i,y_i) \right) \cdot S(\mathbf{x}, \mathbf{y})
\]

Model Parameters: feature weights constitute the set of model parameters, some of which are fixed and some are free to be estimated. More free parameters make the CRF model more complex, which might be harder to learn. The set of free parameters are modeled to avoid redundant parameters, which will not make any contribution. Also, parameters that are not likely to be properly estimated from training data should never be included, because including them will only increase the chance of overfitting the model. Our focus is on the weight of state transition features, because they account for a large proportion of the whole parameter set and good estimation of the weights are critical for successfully predicting TFBSs. A detailed analysis is presented in Supplementary Material.

In the CRF model, we assign a parameter as a weight to each of the features defined previously which are collectively the vector \( \lambda \) in Equation (1). Not all of these parameters are free parameters. Among state transition parameters, we constrain an M state to be only directly reachable from a C state, and not from a G state, since motifs are not present outside CRMs. Thus, state transition features corresponding to taboo transitions have a weight \(-\infty \) (a low enough number in practice), meaning that the transitions never occur in the CRM model. However, we want to have a reasonable number of free model parameters as more free parameters increase the expressibility of the model. With increase in the number of free parameters, the hardness of estimating model parameters increase, the running time of the learning algorithm also rises and some parameters may overfit due to data scarcity for corresponding features.
Alternatively, we make a positive prediction whenever the marginal probability is above a threshold, known as threshold decoding. It is a good baseline measure for the CRF model in motif prediction, critical for improving predictability. The PWM score provides a good power of the PWM constraints feature varies from PWM to PWM. Another feature score tends to suffer from noisy data. Because of this, the discriminative PWM constraints feature vary from PWM to PWM. For PWMs with poor discriminative power, additional features are critical for improving predictability. The PWM score provides a good baseline measure for the CRF model in motif prediction, though it is not an essential feature in our model.

State transition: state transition features are an effort to model the architecture of the regulatory region. The state transition feature
models the relationship between the functionality of neighboring nucleotides, which correspond to neighboring states in the CRF and is based on the differing likelihoods of the hidden CRF states transitioning from one to the other. Details of the mathematical modeling of this feature is provided in Supplementary Material.

Evolutionary conservation and presence or absence of evolutionary events like duplication and repeats can also play a role in identifying TFBS, as evidenced by the large body of work in phylogenetic motif finding. The basic premise in such cases is that functionally relevant nucleotides like TFBS would be under selection, and would hence be distinguishable from surrounding sequence on the basis of evolutionary parameters. While we do not explicitly use multiple species sequence data, we implicitly use evolutionary data in terms of feature data.

**Presence of repeats:** Interspersed repeats and low complexity DNA sequences are common elements in the genome, often near coding regions and inside regulatory sequences. The repeat feature is a simple single nucleotide-based feature indicative of whether that nucleotide is part of a repeat as predicted by RepeatMasker using the repeat database RepBase (Jurka et al., 2005). On one hand, repeats with motif-like patterns may lead to a large number of FP results, but repeats have also been reported to have been under purifying selection (Britten, 1994) and to have been harnessed into the regulatory machinery (Kamal et al., 2006). Thus, instead of masking out repeats to lower the FP rate, we choose to identify repeats in the sequence in a bid to find locational correlations with TFBSs.

**PhastCons score and related features:** We use the PhastCons score as an evolutionary score-based feature. PhastCons (Margulies et al., 2003) is a phylogenetic 2-state HMM which predicts if nucleotide positions in a multiple alignment are in an evolutionarily conserved state or not. The PhastCons score at a nucleotide position is merely the posterior probability that the nucleotide was generated from the conserved state based on the 15-way Multiz (Blanchette et al., 2004) alignment of the Drosophila species, Anceleus gambiae and Tribolium castaneum. We also use two other derived binary features which we feel to be discriminative based on an empirical analysis of PhastCons score distributions (Fig. 3): ‘Is PhastCons score <0.05’ and ‘Is PhastCons score >0.95’. We also keep an additional feature indicating whether PhastCons data are available or not for bookkeeping purposes.

It is well established in the literature that the distance of the TFBS to the TSS plays an important role of the efficacy of the TFBS in regulating the gene (Defrance and Touzet, 2006; Kim et al., 2008; Tharakaraman et al., 2005), and of the nature of function of the TFBS (Elmskis et al., 2006). We therefore incorporate several features which contain information of the distance to the TSS, the locations of the transcribed and translated regions, and the positioning of binding site with respect to the gene transcription-translation direction.

**Distance to TSS and translated:** TFBS are typically present near coding sequences, and we utilize two features indicative of that fact. The binary feature ‘Translated’ indicates at each nucleotide position whether it is translated or not by the gene translation/transcription machinery. It has also been shown that TFBSs are not uniformly distributed wrt their distance from the TSS (Defrance and Touzet, 2006), and the Distance to TSS feature is a score of the distance of each nucleotide from the TSS in question.

**5’UTR and 3’UTR:** The position of the TFBS wrt directionality of the gene being coded has been shown to be a discriminative feature for identifying TFBS. We use two binary features indicative of this fact, the ‘5’UTR’ feature indicates for each nucleotide if it is located in the 5’ untranslated region, and the ‘3’UTR’ feature indicates likewise for the 3’ untranslated region.

Recent work in the literature has approached the TFBS prediction problem as a non-binary classification problem, instead choosing to model the affinity of a TF to bind to a particular oligonucleotide sequence with an affinity score (Ward and Bussemaker, 2008). This has led to the realization that TFBSs may also be effective gene regulators in cases of low binding affinity but high chromatin stability and accessibility (Ozsolak et al., 2007). While we model our TFBS prediction as a sort of classification problem, we still incorporate the notions of chromatin accessibility and stability.

**GC content and melting temperature:** The GC content feature of a genomic sequence or the fraction of G+C bases in a sequence is a simple heuristic which can be used to estimate several factors reflective of the stability of the chromatin structure like the melting temperature and in higher eukaryotes is a determining factor for
identifying CpG islands (Zhang, 2007), thus being indicative of how easy it might be for a TF to actually bind in the locality. The window size \( w \) for the genomic neighborhood over which to estimate the GC content is a hyperparameter that must be determined ahead of time, and is usually chosen to be of the order of magnitude of the binding site. The melting temperature feature is defined as the temperature for which half the DNA strands of an oligonucleotide are in the double helical structure, while the other half are in a random coil formation. It corresponds strongly to chromatin stability, and has been shown as a feature to correlate well with TFBS (Ponomarenko et al., 1999).

**Nucleosome occupancy:** Recent research has suggested that nucleosome occupancy has a strong correlation with binding preference of TFs (Segal et al., 2006). This is due to the non-feasibility of access to the chromatin by the TF when a nucleosome is already bound there. Some research has successfully used nucleosome occupancy scores to improve TFBS predictions (Naflikari et al., 2007).

We also tried several other features directly computable from sequence information, and found that the following features can help in discriminating between TFBS and non-TFBS. The cause of the discriminative power of these tracks may stem from the nature of the binding specificities of the TFs in question, and a closer investigation is warranted.

**Reverse complementarity and conservation symmetry:** We also try two additional features for the CRF based on symmetry of the oligonucleotide in question. The reverse complementarity feature indicates as a fraction between 0 and 1 how similar a nucleotide sequence is to its reverse complement. It is exactly 1 only when an oligonucleotide sequence is identical to its reverse complement. The conservation symmetry feature models how symmetric the degree of conservation in the PWM is wrt the center of the binding site. This is based on the empirical observation that DNA binding domain binding specificities often have symmetric sequence conservation profiles.

The design of new features has exciting new possibilities. Long-range regulatory effects have been reported in the literature (Carroll et al., 2005). The CRF model also readily enables us to model long-range dependencies if we deviate from the chain structured CRF structure. It can also be used as a form of ensemble learning by incorporating predictions by other independent tools as features. Other features which have been shown in the literature to correlate well with the data and which are candidates for future inclusion on this and other datasets include the presence of the nucleotide in the first intron of the regulated gene, and presence of the nucleotide in the neighborhood of a CpG island.

We tested the discriminative nature of these features on the dataset in Figure 3. Figure 3a shows the difference in mean values for background, CRM and motif nucleotides for two of the most discriminative features: GC content and PhastCons score. Figure 3b shows the distribution of PhastCons scores in motifs versus non-motif nucleotides, with the most discriminative bins being at either end of the score range, which offered us some insight as to how to define a derived feature which is more discriminative than the original one. Figure 3c shows the interesting multimodal distribution of the normalized and transformed values of the feature distance to the TSS, suggesting a complicated, non-uniform distribution worth additional investigation.

**3.2 Experimental setup**

In this part, we include biological and empirical bases for selection of some features, data preparation, hyper-parameter setting, test scheme and evaluation scheme. For training data, we use a part of the sequences with ground truth labels. For testing, the required hyper-parameters in the CRF model are the window size used in GC percentage calculation and pseudo-counts used to smooth the probabilities in PWMs to allow for greater tolerance in motif discovery. We set the window size of GC percentage to 8 bps (approximately the average length of a motif) and pseudo-count for smoothing PWM probabilities to 0.5.

Our evaluation is based on a leave-one-out cross-validation (LOOCV) scheme. Each time we take all but one sequences as training data, and predict on the remaining sequence by the model with parameters learnt from the training data. We use the rank decoding scheme with the prediction factor \( k \) set to 0.0015 by

---

**Fig. 3.** (a) Means of two discriminative features plotted for GC content and PhastCons score for Motifs, CRMs and background nucleotides. (b) distribution of PhastCons scores in motifs versus non-motifs and (c) multimodal empirical distribution of feature values for the transformed distance to TSS feature.
default. This threshold is obtained by analyzing the empirical density of TFBSs in training data. Varying the value of the threshold results in increasing one of the performance metrics of precision (P) or recall (R) at the cost of the other. For evaluating performance, we use the standard definitions of P, R and the F1 score using counts of TP, FP and false negative (FN) prediction instances. The exact method of calculating the evaluation metrics is given in Supplementary Material.

Specificity scores and ROC curves are not shown as these evaluation schemes are inappropriate in the context of motif detection. True negative (TN) instances in ground truth may be discovered to contain motifs in the future. Also, the number of positive instances and number of predictions are much smaller than the number of total instances, causing the specificity to be very close to 1 almost always.

3.3 Tests on features

We have empirically established the discriminative nature of our feature set, but we also examine the soundness of the designed features in the context of the CRF model after incorporating some basic features, before including all of them in the model to test for feature redundancy and compatibility in the CRF framework. The state transition features and sequence conservation features are fundamental, so we check the validity of the other features based on predictions made by a basic model consisting of only state transition features and sequence conservation features. The soundness of additional feature is shown by comparing the distributions of the set of TPs and the set of FPs as predicted by the basic model.

We learn a CRF model using the two kinds of fundamental features, and use it to get a set of predictions of TFBSs, which contains both TP predictions and FP predictions. We split the predictions into two groups, TP group and FP group, and compute the GC percentage score, reverse complementary score and conservation symmetry score for each of the instances in the two groups. We can show the soundness of a feature by a statistical analysis on the difference between scores of the two groups. There are 193 instances in TP group and 499 instances in FP group. Comparisons of cumulative distribution function (CDF) curves between TP group and FP group on GC percentage scores, reverse complementary scores and conservation symmetry scores are shown in Figure 4. The scores plotted are raw scores without an offset, such as $p$, $s$ and $cs$ in Equations (9), (11) and (13) of Supplementary Material. We can see that the CDF curve of TP group is almost always lower than that of FP group in GC percentage score and reverse complementary score, while the CDF curve of TP group is almost always higher than that of FP group in conservation symmetry score.

For the feature of GC percentage, the scores in TP group have a mean at 0.4641 and sample variance at 0.0043, and the scores in FP group have a mean at 0.4323 and sample variance at 0.0065. Assuming that they both follow Gaussian distributions, we have a difference between means at 0.0318 with a SD at 0.0059, which gives us a confidence value at $1 - 4 \times 10^{-8}$ that the mean of TP group is bigger than the mean of FP group. It is credible that GC percentage feature is informative. Following a similar analysis, for the feature of reverse complementarity, the mean TP score is 0.3041 and sample variance 0.0349, and the mean FP score is 0.2413 and sample variance 0.0360. With a difference between means at 0.0159 with a SD at 0.0059, we have a confidence value at $1 - 4 \times 10^{-5}$ that the mean of TP group is bigger than the mean of FP group. For the feature of conservation symmetry, the TP scores have mean 0.5215 and sample variance 0.0541, and the FP scores have a mean 0.5950 and sample variance 0.0666. The confidence value that TP group has a smaller average score than FP group is $1 - 1.5 \times 10^{-4}$.

3.4 Performances on TFBS prediction

Synthetic dataset: We compare the CRF model with BayCis, ClusterBuster and Cister on the synthetic TRS dataset. CRF model outperforms ClusterBuster and Cister but not BayCis (Fig. 5a) on the synthetic dataset. BayCis has an advantage over the other tools
Drosophila we outperform Baycis on the real dataset. Having the same background model as the simulation scheme, but in comparison with other algorithms at their default settings on the real synthetic dataset, (Fig. 5. (a) P–R performance of CRF, BayCis, Cluster-Buster and Cister on the synthetic dataset, (b) F1 score and (c) P–R curve of the CRF model in comparison with other algorithms at their default settings on the real D. melanogaster TRS dataset.

Fig. 5. (a) P–R performance of CRF, BayCis, Cluster-Buster and Cister on the synthetic dataset, (b) F1 score and (c) P–R curve of the CRF model in comparison with other algorithms at their default settings on the real D. melanogaster TRS dataset.

ACKNOWLEDGEMENTS

The authors thank Geir Kjetil Sandve and Veronica Hinman for comments and suggestions.

Funding: National Science Foundation (CAREER Award grant DBI-0546594 to E.P.X.); Alfred P. Sloan Research Fellowship (to E.P.X.).

Conflict of Interest: none declared.

REFERENCES

Alkema, W.B. et al. (2004) Mscan: identification of functional clusters of transcription factor binding sites. Nucleic Acids Res., 32, W195–W199.

Avriel, M. (2003) Nonlinear Programming: Analysis and Methods. Dover Publishing, Mineola, NY.

Berman, B.P. et al. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl Acad. Sci. USA, 99, 757–762.

Blanchette, M. et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res., 14, 708–715.

Bochkov, N. and Craven, M. (2004) Markov networks for detecting overlapping elements in sequence data. Proc. Adv. Neural Inform. Process. Syst., 17, 193–200.

Boyd, S. and Vandenberghe, L. (2004) Convex Optimization. Cambridge University Press, Cambridge.

Botten, R. (1994) Evolutionary selection against change in many Alu repeat sequences interspersed through primate genomes. Proc. Natl Acad. Sci. USA, 91, 5902–5906.

Bulyk, M. et al. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res., 30, 1255–1261.
Carr,B.L. et al. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell, 122, 33–43.

Dumais,T. and Girolami,M.A. (2008) Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. Bioinformatics, 24, 1264–1270.

Dawid,PL.E.H. (2001) Genome Regulatory Systems: Academic Press, San Diego, CA.

DeCristofaro,P. et al. (2007) Correlate gene prediction using conditional random fields. Genome Res., 17, 1389–1398.

DevScan,M. and Touze,D.B. (2008) Predicting transcription factor binding sites using local over-representation and comparative genomics. BMC Bioinformatics, 9, 396.

Donaldson,I.J. et al. (2005) TchiCluster: a resource for the characterization of transcriptional regulatory networks. Bioinformatics, 21, 3056–3059.

Elinson,R.L. et al. (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res., 16, 1455–1466.

Ernst,J. (2008) Computational Methods for Analyzing and Modeling Gene Regulation Dynamics. PhD dissertation, Carnegie Mellon University, ML.

Frid,H.M.C. et al. (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. Nuclear Acids Res., 30, 3214–3224.

Frid,H.M.C. et al. (2005) Cluster-buster: finding dense clusters of motifs in DNA sequences. Nuclear Acids Res., 33, 3666–3668.

Galilo,S.M. et al. (2006) Redfly: a regulatory element database for drosophila. Bioinformatics, 22, 361–363.

Girod,K. et al. (2007) CONTRAST: a discriminative, phylogeny-free approach to multiple informatic methods for multiple gene prediction. Genome Biol., 8, R269.

Johansson,P. et al. (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the mscan algorithm. Bioinformatics, 19 (Suppl. 1), i169–i176.

Kulakova,T. et al. (2005) RepBase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res., 108, 462–467.

Kumar,M. et al. (2006) A large family of ancient repeat elements in the human genome is under strong selection. Proc. Natl Acad. Sci. USA, 103, 2740–2745.

Kim,N.K. et al. (2008) Finding position motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites. BMC Bioinformatics, 9, 262.

Laflatte,J. et al. (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In Proceedings of the 1st International Conference on Machine Learning (ICML), 2001, Williamsburg, MA.

Lin,T-H. et al. (2008) Baycius: a bayesian hierarchical lmm for cis-regulatory module detection in metazoan genomes. In Proceedings of RECOMB 2008, Singapore.

Louie,G.G. et al. (2002) Visota for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res., 12, 832–839.

Marques,E.H. et al. (2003) Identification & characterization of multi-species conserved sequences. Genome Res., 13, 2507–2518.

Michelson,A.M. (2002) Deciphering genetic regulatory codes: a challenge for functional genomics. Proc. Natl Acad. Sci. USA, 99, 546–548.

Moses,A.M. et al. (2004) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. In Proceedings of Pacific Symp. Biocomput. 2004, Hawaii, pp. 324–335.

Nurung,V. et al. (2006) Computational annotation of transcription factor binding sites in D. melanogaster developmental genes. In Proceedings of The 17th International Conference on Genome Informatics. Yokohama.

Nurilaki,L. et al. (2007) A nucleosome-guided map of transcription factor binding sites in yeast. PLoS Comput. Biol., 3, e215.

Naughton,B. et al. (2006) A graph-based motif detection algorithm models complex nucleotide dependencies in transcription factor binding sites. Nucleic Acids Res., 34, 576–5739.

Nota,K. and Coven,M. (2007) Learning probabilistic models of cis-regulatory modules that represent logical and spatial aspects. Bioinformatics, 23, e156–e162.

Otsuka,M. et al. (2007) High-throughput mapping of the chromatin structure of human promoters. Nat. Biotechnol., 25, 244–248.

Pearl,J. (1988) Probabilistic Reasoning in Intelligent System: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA.

Ponomarev,J. et al. (1999) Conformational and physical-chemical DNA features specific for transcription factor binding sites. Bioinformatics, 15, 654–668.

Pudim,R. et al. (2006) Feature based representation and detection of transcription factor binding sites. In Proceedings of the German Conference on Bioinformatics 2006. Bonn, pp. 43–52.

Regev,A. et al. (2002) Computational detection of genomic cis-regulatory modules applied to body patterning in the early drosophila embryo. BMC bioinformatics, 3, 30.

Ray,J. et al. (2009) Comet: comparative genomic motif detection via multi-resolution phylogenetic shadowing. PLoS Comput. Biol., 5, e1000090.

Rebeczi,C. et al. (2002) Score: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. site clustering over random expectation. Proc. Natl Acad. Sci. USA, 99, 8988–8993.

Sandve,G.K. and Drablos,F. (2006) A survey of motif discovery methods in an integrated framework. Bio. Direct., 1.

Segal,E. et al. (2006) A genomic code for nucleosome positioning. Nature, 442, 722–727.

She,F. and Pereira,F. (2003) Shallow parsing with conditional random fields. Proc. Hum. Lang. Tech-NAACL, 1, 134–141.

Sharan,R. et al. (2003) Creme: a framework for identifying cis-regulatory modules in human-mouse conserved segments. Bioinformatics, 19 (Suppl. 1), 985–991.

Sharan,E. and Segal,E. (2007) A feature-based approach to modeling protein-DNA interactions. Lect. Notes Comput. Sci., 4453, 77–91.

Siddharthan,R. et al. (2004) PhyloGibbs: a gibbs sampler incorporating phylogenetic information. In Eskin,E. and Workman,C. (eds), Regulatory Genomics, Vol. 3318 of Lecture Notes in Computer Science, Springer, Berlin, pp. 30–41.

Simha,S. and He,X. (2007) MORPH: probabilistic alignment combined with hidden Markov models of cis-regulatory modules. PLoS Comput. Biol., 3, e216.

Simha,S. et al. (2004) PhyMe: a probabilistic algorithm for finding motifs in sets of orthologous sequences. BMC bioinformatics, 5, 170.

Simha,S. et al. (2006) Stab: a program for discovery and analysis of cis-regulatory modules. Nuclear Acids Res., 34, W555–W559.

Simha,S. et al. (2007) Systematic functional characterization of cis-regulatory motifs in human core promoters. Genome Res., 18, 477–488.

Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. Nuclear Acids Res., 12, 505–519.

Tharakaram,K. et al. (2005) Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. Bioinformatics, 21 (Suppl. 1), i316–i321.

Thompson,W. et al. (2004) Decoding human regulatory circuits. Genome Res., 14, 1967–1974.

Ward,L. and Bussemaker,H. (2008) Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. Bioinformatics, 24, 1685–1711.

Wingender,E. et al. (2000) TRANSFAC, an integrated system for gene expression regulation. Nuclear Acids Res., 28, 316–319.

Zhang, M. (2007) Computational analyses of eukaryotic promoters. BMC Bioinformatics, 8 (Suppl. 6), S3.