BIOINFORMATICS ORIGINAL PAPER

Vol. 22 no. 2 2006, pages 157–163
doi:10.1093/bioinformatics/bti731

Sequence analysis

Sequence features of DNA binding sites reveal structural class of associated transcription factor

Leelavati Narlikar and Alexander J. Hartemink*

Duke University, Department of Computer Science, Box 90129, Durham, NC 27708, USA

Received on July 21, 2005; revised on September 30, 2005; accepted on October 18, 2005
Advance Access publication November 2, 2005
Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: A key goal in molecular biology is to understand the mechanisms by which a cell regulates the transcription of its genes. One important aspect of this transcriptional regulation is the binding of transcription factors (TFs) to their specific *cis*-regulatory counterparts on the DNA. TFs recognize and bind their DNA counterparts according to the structure of their DNA-binding domains (e.g. zinc finger, leucine zipper, homeodomain). The structure of these domains can be used as a basis for grouping TFs into classes. Although the structure of DNA-binding domains varies widely across TFs generally, the TFs within a particular class bind to DNA in a similar fashion, suggesting the existence of class-specific features in the DNA sequences bound by each class of TFs.

Results: In this paper, we apply a sparse Bayesian learning algorithm to identify a small set of class-specific features in the DNA sequences bound by different classes of TFs; the algorithm simultaneously learns a true multi-class classifier that uses these features to predict the DNA-binding domain of the TF that recognizes a particular set of DNA sequences. We train our algorithm on the six largest classes in TRANSFAC, comprising a total of 587 TFs. We learn a six-class classifier for this training set that achieves 87% leave-one-out cross-validation accuracy. We also identify features within *cis*-regulatory sequences that are highly specific to each class of TF, which has significant implications for how TF binding sites should be modeled for the purpose of motif discovery.

Contact: lee@cs.duke.edu; amink@cs.duke.edu

1 INTRODUCTION

Transcriptional regulation of a eukaryotic gene is governed in large part by interactions between molecules called transcription factors (TFs) and their corresponding binding sites on the DNA near the regulated gene. A particular TF often recognizes and binds to multiple different nucleotide sequences; only rarely is the sequence that a TF recognizes non-degenerate. While there is typically some degeneracy in the sequences of TF binding sites, enough similarity apparently exists in the sequences to ensure sufficient specificity of TF binding.

Many different models have been posited for representing the similarities within the set of sequences bound by a particular TF. The most commonly used probabilistic representation is the position-specific scoring matrix (PSSM; also called a position weight matrix, PWM), which stores the preference for each putative nucleotide at each position of the binding site (Staden, 1984). The PSSM model assumes (1) that the binding sites recognized by a particular TF are of a fixed length and (2) that position-specific nucleotide preferences exhibit independence between positions. The latter assumption has been shown to be untenable in certain cases (Bulyk *et al*., 2002), and more flexible probabilistic representations like weight array matrices with a first order Markov dependence assumption (Zhang and Marr, 1993), trees or Bayesian networks (Agarwal and Bafna, 1998), and mixtures of trees or PSSMs (Barash *et al*., 2003) relax this latter assumption. However, all these models still retain the former assumption, namely that the binding sites for a particular TF are of a fixed length. In contrast, non-probabilistic models typically represent the degeneracy within binding sites for a particular TF using regular expressions, which often permit variable length sites to be represented. A model that incorporates the robustness of a probabilistic framework with the ability to represent variable length sites would seem to be ideal, but something further is also desirable. Since different classes of TFs bind to DNA differently, an ideal model should represent the binding sites of different classes of TFs differently.

TFs are generally proteins. Hence, they have different three-dimensional structures and in particular, different domains for recognizing and binding to DNA. The binding affinity of a TF depends on both its DNA-binding domain and the specific sequence of nucleotides on the DNA that is recognized. For instance, a protein from the leucine zipper family that binds as a homodimer will be more likely to recognize a palindromic binding site than a homeodomain protein, which is known to be a stable monomer and has no mirror symmetry in its shape. In this paper, we identify strong predictive relationships between the three-dimensional DNA-binding domain of various classes of TFs and the sites that each class tends to recognize. We do this by building a sparse Bayesian multi-class classifier that can accurately classify the binding domain of a TF based only on features that are present in the DNA sequences to which the TF binds, achieving 87% leave-one-out-cross-validation (LOOCV) accuracy. This indicates the presence of features that are characteristic to the sequences recognized by different classes of proteins.

Another application of our work is to help identify the specific TF that is responsible for regulating a set of co-regulated genes. Conventional motif finding algorithms can be applied to the upstream regions of genes that are suspected to be co-regulated, on the basis of gene expression data for instance. Although a common conserved sequence motif is often found, the identity of the corresponding TF remains unknown. Knowing the TF's DNA-binding domain can

*To whom correspondence should be addressed.

help narrow the search. Here, we demonstrate how our classifier can be applied to motifs detected by conventional motif finding algorithms to predict the DNA-binding domain with high accuracy.

We stress that the immediate purpose of this paper is neither to propose a new model for representing binding sites, nor to develop an algorithm for finding them, but rather to show that strong predictive relationships exist between sets of binding sites and the classes of TFs that bind them. Nevertheless, our results have larger implications in that they suggest a more expressive model for representing binding sites is required, and that new algorithms for finding them should be developed. We return to these points in the discussion.

2 TRANSCRIPTION FACTOR CLASSES

To identify relationships between the structure of the DNA-binding domain of a TF and the corresponding DNA sequences that it recognizes, we first need to separate the TFs into classes so that TFs with similar DNA-binding domains are grouped together. For this purpose, we use the TRANSFAC 6.0 database (Wingender *et al*., 2001) which maintains a comprehensive list of eukaryotic TFs, along with their structural domains and the sequences to which they are known to bind. TRANSFAC organizes TFs in a hierarchy; at the top level are four superclasses: zinc-coordinating, basic domains, helix–turn–helix (HTH) and beta scaffold. Each superclass is further divided into a large number of classes (which in turn are divided into families and subfamilies). In this paper, we concentrate on the six biggest TRANSFAC classes which, incidentally, include representatives from three of the four superclasses: Cys_2His_2 zinc fingers and Cys_4 zinc fingers (both within the zinc-coordinating superclass); basic helix–loop–helix (HLH) and basic leucine zippers (both within the basic domain superclass); and forkheads and homeodomains (both within the HTH superclass). We use these six classes as the classes for our classifier. The characteristics and binding mechanisms of each of the six classes are described in more detail in the following six sections.

2.1 Class I: Cys_2His_2

Like other zinc-coordinating TFs, Cys_2His_2 proteins contain a self-folding DNA-binding domain in which zinc is a crucial component for the stability of its tertiary structure. This binding domain is called a zinc finger, and each finger has an α-helix and a β-sheet held together by the zinc ion. Usually a TF belonging to this class contains multiple such fingers, with each finger making contact with the major groove of DNA in series (Fig. 1A). The name of the class follows from the fact that each zinc finger contains 2 cysteine and 2 histidine residues that coordinate one zinc ion.

2.2 Class II: Cys_4

Like Class I, Cys_4 proteins also belong to the zinc-coordinating superclass of TFs and are often nuclear receptors. The TFs in this class generally bind to DNA before binding to a hormone, helping the cell detect the presence of and alter transcription in response to the respective hormone. The DNA-binding domain of a Cys_4 TF contains two zinc finger motifs differing in size, composition and function. Each finger contains four cysteine residues that coordinate one zinc ion. The helix of the first finger binds to the major groove of the DNA, while the sequence of the second finger mediates dimerization after DNA binding. By comparing Figure 1A and 1B, we see that although both Cys_2His_2 and Cys_4 TFs have a zinc atom as an important constituent, they have different structures and binding mechanisms.

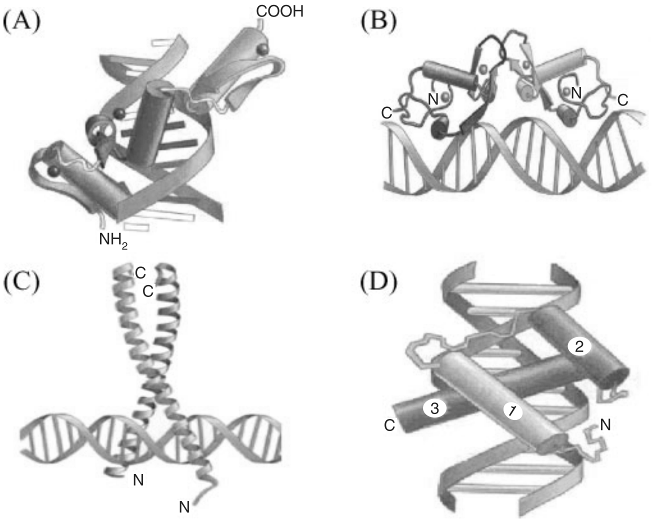

Fig. 1. Different classes of TFs bind DNA differently. (**A**) A mouse regulatory protein composed of different zinc fingers, a member of the Cys_2His_2 class of zinc-coordinating TFs. (**B**) An adrenal gland glucocorticoid receptor, a dimer from the Cys_4 class of zinc-coordinating TFs. (**C**) A dimer of yeast Gcn4, a member of the basic leucine zipper class. (**D**) A yeast regulatory protein, MATα2, belonging to the homeodomain class of TFs. (Adapted from Alberts *et al*., 2002.)

2.3 Class III: bHLH

Basic helix–loop–helix (bHLH) proteins belong to the superclass of basic domains. As the name suggests, TFs in this superclass possess a basic region that makes contact with the DNA. In bHLH proteins, the basic region is followed by a HLH motif consisting of a short α-helix connected by a loop to a second, longer α-helix. The flexibility of the loop allows one helix to fold back and pack against the other. This HLH motif mediates dimerization as a prerequisite for DNA-binding. The DNA-binding specificity is dictated by the basic region, which adopts an α-helical conformation upon binding.

2.4 Class IV: bZip

Basic leucine zipper proteins also belong to the superclass of basic domains and possess a basic DNA-binding region. They exhibit the unique trademark of having a leucine residue at every seventh position along a long α-helix over a distance covering eight helical turns. The periodically spaced leucine residues are directed toward one side of the helix. The leucine side chains extending from the α-helix of one bZip TF interact with those in the α-helix of a second bZip as shown in Figure 1C, facilitating dimerization. As can been seen from the figure, the upper parts are zipped together by leucine residues (not explicitly displayed) and the lower parts, which are basic in nature, grip the DNA molecule like a clothespin on a clothesline. As with the bHLH class of TFs, the basic regions adopt an α-helical conformation upon binding.

2.5 Class V: forkhead

Forkhead proteins belong to the superclass of HTH domain TFs. Originally found in bacterial proteins, HTH domains consist of two

α-helices: one lies in the wide groove of the DNA and the other lies at an angle across the DNA. These are not to be confused with HLH domains, which have a characteristic linker region of varying length between their two α-helices and are more prone to dimerization. Forkhead TFs contain 100 amino acids that are evolutionarily well conserved and essential for DNA recognition, and usually bind DNA as monomers (Aitola, 2002). The forkhead domain includes two loops, or wings, on the C-terminal side of the HTH domain giving it a butterfly-shaped structure (also called a 'winged helix motif').

2.6 Class VI: homeodomain

Homeodomain proteins belong to the same HTH superclass as forkhead proteins. The homeodomain is composed of three α-helices that are packed together by hydrophobic interactions. The latter two helices comprise the HTH motif. The recognition helix (helix 3 in Fig. 1D) makes important contacts with the major groove of the DNA. Nucleotide pairs are also contacted in the minor groove by a flexible arm attached to the first helix.

3 METHODS

Our goal is to produce a true multi-class classifier that is capable of predicting the type of DNA-binding domain for a TF based only on features that are present in the DNA sequences to which the TF binds. To this end, we formulate a set of possibly relevant sequence features, and then apply a sparse Bayesian learning algorithm to identify a small subset of truly relevant features within the set of possibly relevant features. The algorithm simultaneously learns a multi-class classifier that uses this subset of truly relevant features to predict the type of DNA-binding domain for a given TF. In the sections that follow, we describe the sequence features we extracted, the sparse Bayesian learning algorithm we applied, and the training set we used to learn our multi-class classifier.

3.1 Sequence features

For each TF, we are given a set of DNA sequences to which it binds, and for each DNA binding site in this set, we construct a vector of length 1387 describing possibly relevant sequence features present in the binding site. These features include the following:

(1) Subsequence frequency features (1364): Integers representing counts of all subsequences of length 1 (i.e. each of the 4 nt) to length 5 (i.e. each of the 4^5 possible nucleotide strings). These integers account for a total of 1364 entries in the vector, comprising the vast majority of possibly relevant features.

(2) Ungapped palindrome features (8): Binary indicator variables denoting whether the binding site contains palindromic subsequences of half-length 3, 4, 5 or 6 that span the entire site (i.e. end to end), as well as those that do not span the entire site (i.e. are somewhere in the middle of the site).

(3) Gapped palindrome features (8): Binary indicator variables denoting whether the binding site contains gapped palindromic subsequences of half-length 3, 4, 5 or 6 that span the entire site (i.e. end to end), as well as those that do not span the entire site (i.e. are somewhere in the middle of the site). A gapped palindromic subsequence is one in which some non-palindromic nucleotides are inserted exactly in the middle of two otherwise palindromic halves.

Table 1. Special sequence features

Sequence feature	Class name	Reference
`G..G`	Cys_2His_2	Adapted from Wolfe *et al*. (2000)
`G..G..G`	Cys_2His_2	Adapted from Wolfe *et al*. (2000)
`[GC]..[GC]..[GC]`	Cys_2His_2	Adapted from Wolfe *et al*. (2000)
`AGGTCA | TGACCT`	Cys_4	Zilliacus *et al*. (1995)
`CA..TG`	bHLH	Atchley and Fitch (1997)
`TGA.*TCA`	bZip	Mulder *et al*. (2003)
`TAAT | ATTA`	Homeodomain	Pabo and Sauer (1992)

Special sequence features that have been identified in the literature to be over-represented in the binding sites of certain classes of TFs. All features in the table are formatted as regular expressions, where standard conventions apply.

(4) Special features (7): Binary indicator variables that denote the presence or absence of features that have been identified in the literature to be over-represented in the binding sites of certain classes of TFs. The seven features we used are listed in Table 1.

We then construct a composite feature vector for each TF by computing the arithmetic average of the feature vectors for each DNA binding site bound by the TF. To this composite vector, we append three binary indicator variables denoting whether the TF is found in animal, plant or fungus, resulting in a final feature vector of length 1390.

We have intentionally constructed this feature vector to be larger than we might otherwise because our learning algorithm is explicitly designed to utilize only a small subset of these features when learning a classifier. More technically, because the learning algorithm is regularized by a sparsity-promoting l_1 prior, it tries to use only those features which provide strong predictive value and no more, so as to prevent over-fitting (the ability of sparse Bayesian learning algorithms to accomplish this goal has been discussed in a large number of papers, including Williams (1995); a summary is provided in Krishnapuram *et al*. (2005)). As a result, omitting features that might be relevant to the prediction is of greater concern than including features that might be irrelevant. As we demonstrate below, the feature vector we construct includes sufficient information to permit high classification accuracy across six different classes of TF, which suggests that a subset of the features is indeed relevant. That said, there may still exist other features that would improve either the accuracy or the sparsity of the learned classifier if they were to be included.

3.2 Sparse Bayesian learning algorithm

We anticipate that most of the features in our feature vectors will be irrelevant for distinguishing between the six classes of TFs. This has the potential to hurt the generalization performance of a classifier built on these features unless the classifier is specifically designed to guard against over-fitting. To this end, we have chosen to apply the sparse multinomial logistic regression (SMLR) algorithm of Krishnapuram *et al*. (2005).

The SMLR algorithm learns a true multi-class classifier and simultaneously performs feature selection to identify a small subset of features relevant to the class distinctions. The learned classifier reports the probability of a sample belonging to each of the m classes given m sets of feature weights, one for each class.

In particular, if $\boldsymbol{y} = [y^{(1)}, \ldots, y^{(m)}]^{\mathrm{T}}$ is a 1-of-m encoding of the m classes, and if $\boldsymbol{w}^{(i)}$ is the feature weight vector associated with class i, then the probability that a given sample $\boldsymbol{x}$ belongs to class i is given by

$$P(y^{(i)} = 1|\boldsymbol{x}, \boldsymbol{w}) = \frac{\exp(\boldsymbol{w}^{(i)^{\mathrm{T}}}\boldsymbol{x})}{\sum_{i=1}^{m}\exp(\boldsymbol{w}^{(i)^{\mathrm{T}}}\boldsymbol{x})},$$

where $\boldsymbol{w} = [\boldsymbol{w}^{(1)^{\mathrm{T}}}, \ldots, \boldsymbol{w}^{(m)^{\mathrm{T}}}]^{\mathrm{T}}$.

The vector $\boldsymbol{w}$ is said to be sparse if and only if many of its entries are exactly zero, which implies that many of the features in the feature vector are irrelevant in making a class prediction. In order to achieve sparsity in our estimate of $\boldsymbol{w}$, we incorporate a sparsity-promoting l_1 prior on the entries of $\boldsymbol{w}$, and then estimate $\boldsymbol{w}$ from the data using a maximum a posteriori (MAP) criterion in place of the typical maximal likelihood (ML) criterion for multinomial logistic regression. Explicitly, we compute

$$\hat{\boldsymbol{w}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{w}}[\ell(\boldsymbol{w}) + \log p(\boldsymbol{w})],$$

where $p(\boldsymbol{w})$ is a prior distribution on $\mathbf{w}$ and $\ell(\boldsymbol{w})$ is the log-likelihood function:

$$\ell(\boldsymbol{w}) = \sum_{j=1}^{n}\left[\sum_{i=1}^{m} y_j^{(i)}\boldsymbol{w}^{(i)^{\mathrm{T}}}\boldsymbol{x}_j - \log\left(\sum_{i=1}^{m}\exp\left(\boldsymbol{w}^{(i)^{\mathrm{T}}}\boldsymbol{x}_j\right)\right)\right],$$

where n is the number of training samples.

We choose a sparsity-promoting Laplacian prior on the parameters, which means that $p(\mathbf{w}) \propto \exp(-\lambda\|\boldsymbol{w}\|_1)$ where λ acts as a tunable regularization parameter. The reason for using a Laplacian prior is that its l_1-norm penalty promotes sparsity, setting many weights to exactly zero (Williams, 1995); the larger is λ, the greater is the sparsity. Excessively large values of λ can result in the non-selection of relevant features, while excessively small values of λ can result in the selection of irrelevant features; in each case generalization performance is expected to suffer, suggesting that cross-validation can be used to optimize λ, which is exactly the procedure we perform. We tried values of 0.01, 0.1, 0.5, 1, 1.5, 2 and 10 for λ and observed a minimum at $\lambda = 1$, so this is the value we choose for the remainder of the paper.

The algorithm for computing $\hat{\boldsymbol{w}}_{\mathrm{MAP}}$ is extremely fast because the objective function is concave; the algorithm exploits a bound optimization framework to perform sequential iterative optimization of an even simpler concave lower bound. Concavity of the objective function also implies that there is a unique global optimum rather than a large number of local optima, so the results are provably optimal and initialization of the iterative optimization procedure is not important. Details of the actual algorithm can be found in Krishnapuram *et al*. (2005).

3.3 Training set

To learn a multi-class classifier, we provide the SMLR algorithm with a training set of $n = 587$ TFs from the $m = 6$ largest classes in TRANSFAC, along with the associated feature vectors $\boldsymbol{x}$ and class labels $\boldsymbol{y}$ for each TF. Table 2 shows the class labels we assigned to the six classes, the number of TFs in each class for which binding site information is available in TRANSFAC, and the number of binding sites for each class. We have a total of 3847 binding sites which make up almost two-thirds of all binding sites listed in TRANSFAC, with an average of 7 sites per TF in our training set.

Table 2. The six classes in the training set

Class label	Class name	No. of TFs	No. of sites
I	$\mathrm{Cys_2His_2}$ (zinc-coordinating)	97	776
II	$\mathrm{Cys_4}$ (zinc-coordinating)	97	734
III	bHLH (basic domain)	61	182
IV	bZip (basic domain)	165	1353
V	Forkhead (helix-turn-helix)	52	281
VI	Homeodomain (helix-turn-helix)	115	621
	Total	587	3847

The six largest classes in TRANSFAC, which are the six classes considered by our classifer. The term in parentheses is the superclass to which each class belongs. The third column indicates the number of TFs in each class for which binding site information is available in TRANSFAC, while the third column lists the total number of binding sites for that class.

Table 3. Confusion matrix for the six classes

	I	II	III	IV	V	VI	Total
I	—	0.03	0.03	0.05	0.04	0.07	0.23
II	0.01	—	0	0.04	0.03	0.01	0.09
III	0.10	0	—	0.02	0	0	0.11
IV	0.03	0.01	0.01	—	0.01	0.02	0.08
V	0.10	0	0	0.04	—	0.04	0.17
VI	0.04	0.02	0.02	0.04	0.03	—	0.15

Entry (i, j) of the confusion matrix indicates the fraction of TFs of class i that were misclassified as class j during LOOCV. The far-right column provides the total misclassification rate for each class.

4 RESULTS

4.1 Classification accuracy

We performed an LOOCV to assess the accuracy of the classifier: a new classifier was learned 587 times by holding one sample out at a time, training on the remaining 586 samples and testing on the one held out. A total of 77 TFs were misclassified during this procedure, indicating an overall accuracy of 87%. Table 3 shows the confusion matrix describing all the misclassifications. Class I ($\mathrm{Cys_2His_2}$) is the most difficult to predict, with an error rate of 23%, while Class IV (bZip) is the easiest, with an error rate of only 8%. We also performed the same LOOCV procedure with the same training set but with all the class labels randomly permuted. With random class labels, we observed an overall accuracy of 20%, as expected (a classifier with no information can achieve at most $165/587 = 28\%$ accuracy on this particular set of TFs). The significant difference between 20% accuracy with random class labels and 87% accuracy with true class labels suggests that our classifier is performing admirably.

We also calculated the training error, which is the number of errors made when predicting the classes of all the TFs in the training set using a classifier learned from the full training set. Only one TF from Class I is labeled as Class IV; all other TFs are classified correctly. This provides further evidence that our classifier is correctly selecting class-specific features.

Ideally, to assess the quality of our classifier, we would compare our results with previously reported results in the same area. The most closely related work of which we are aware is the MultiPrototyper of Xing and Karp (2004) and the familial binding profile-based similarity measure by Sandelin and Wasserman (2004). Although both learn a classifier of a similar nature, their work is different from ours in three important respects:

- We use sequence features from the complete set of binding sites that are bound by a TF; in contrast, they use PSSMs as their input. These PSSMs are not built from the complete set of binding sites because many binding sites do not fit well into a PSSM and are thus ignored by TRANSFAC (in the case of the MultiPrototyper) and JASPAR (in the case of Sandelin and Wasserman (2004)) while building the PSSM. This can happen when the binding sites are of variable length and cannot be represented well by a PSSM or when inclusion of the complete set leads to low-entropy PSSMs. We achieve higher accuracy even without filtering out sequences that fit less well.
- Xing and Karp use TRANSFAC's four superclasses (zinc-coordinating, basic domain, HTH, and beta scaffold) as their target output; in contrast, we use TRANSFAC classes (one level lower in the hierarchy) and choose the six largest as our target output. Sandelin and Wasserman use JASPAR profiles (Sandelin *et al*., 2004) to classify 62 TFs into 11 well-characterized classes. These highly specific classes, restricted to multicellular eukaryotes, do not seem to be a good representation of all eukaryotic TFs.
- They learn a classifier with the goal of discovering new motifs in DNA sequences; in contrast, we learn a classifier with the goal of classifying TFs with great accuracy.

We mention these differences as caveats in comparing our results with those of either of these works. Bearing these caveats in mind, our error rates are significantly lower than the MultiPrototyper, in each case less than half. For the zinc-coordinating superclass, they report an error rate of 37%, while ours is 16%. For the basic domain superclass, they report an error rate of 22%, while ours is 9%. For the HTH superclass, they report an error rate of 38%, while ours is 16%. We cannot compare the performance on the beta scaffold superclass because the six largest TRANSFAC classes that we chose to study do not include representatives from this superclass. Comparison with Sandelin and Wasserman (2004) is not possible due to totally different datasets.

4.2 Feature selection consistency

We closely examined the number of times features were selected as relevant by the SMLR algorithm while learning classifiers during the full LOOCV procedure. Figure 2 shows a histogram of the number of times each feature is selected as relevant over the 587 trials. As can be seen in the figure, most of the weight is concentrated at the two ends, indicating strong consistency in feature selection. Throughout the entire procedure, 1047 features are consistently not selected (they are in <10% of the 587 learned classifiers) and 290 features are consistently selected (they are in >90% of the 587 learned classifiers). Taken together, this means that 96% of the original 1390 features are consistently selected or not selected.

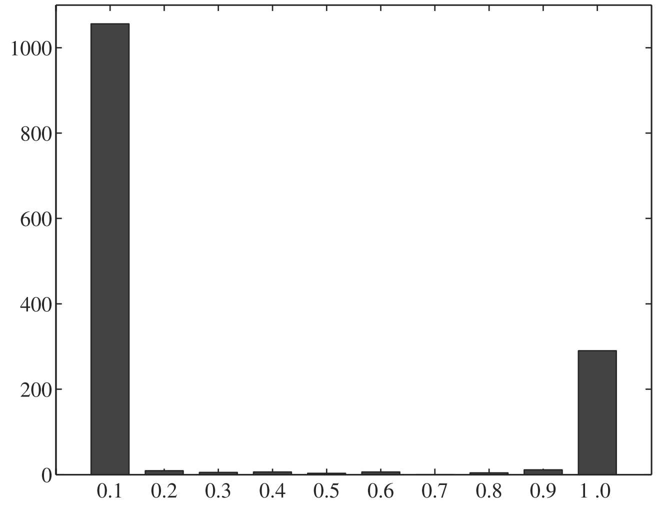

Fig. 2. Histogram depicting the number of times each feature is selected as relevant over the 587 LOOCV trials. The *x*-axis is divided into deciles of the total number of trials. The height of each bar shows the number of features that are selected with that frequency over the LOOCV. Each feature appears exactly once in the graph. The more mass that is concentrated in the first and last bars, the more consistent the classifier is in identifying relevant features.

In particular, 859 features are never selected while 139 are always selected. We repeated the analysis with a 10-fold cross-validation test and found the number of features never selected reduced only slightly to 856 and the number always selected reduced to 103. This further corroborates that the classifier is not over-fitting, but is robustly and consistently identifying relevant features. Some specific details about relevant features are presented in Section 5.

4.3 Supplementing motif finding algorithms

We further tested our classifier to see if it could correctly determine the class of TF binding to motifs discovered by conventional motif finding algorithms. In the context of this paper, the TF classes must already be known for us to evaluate the accuracy of our classifier; outside the context of this paper, our classifier would be applied when the TF classes are unknown. We examined results published by Harbison *et al*. (2004) in which genome-wide chromatin immunoprecipitation (ChIP) experiments are performed to profile the *in vivo* binding of a large number of known yeast TFs. They then use six motif finding algorithms—AlignACE (Roth *et al*., 1998), MEME (Bailey and Elkan, 1994), MDscan (Liu *et al*., 2002), a method by Kellis *et al*. (2003), MEME_c (Harbison *et al*., 2004) and CONVERGE (Harbison *et al*., 2004)—to identify PSSM-based motifs that are present in the genomic sequences bound by each TF with high confidence (P-value < 0.001). Of the 203 TFs profiled by ChIP, Harbison *et al*. report that 65 are associated with a single significant motif (based on these algorithms and additional sequence conservation criteria). Of these 65 TFs, 44 are listed in TRANSFAC along with a domain class, 22 of which belong to one of the six classes in our classifier. Out of these 22 TFs, eight of them were already in our training set which left us with 14 TFs on which to test our classifier. To test, we selected 20 binding sites reported to be associated with each of the 14 TFs, but emphasize that the binding sites are not experimentally confirmed; rather, they result from a

Table 4. Supplementing conventional motif finding algorithms

Number	TF Name	Actual class	Predicted class
1	Azf1	I	I
2	Cad1	IV	IV
3	Cin5	IV	IV
4	Fhl1	V	IV
5	Fkh1	V	V
6	Fkh2	V	V
7	Ino2	III	III
8	Ino4	III	III
9	Met4	IV	IV
10	Nrg1	I	I
11	Phd1	III	IV
12	Sok2	III	IV
13	Tye7	III	III
14	Yap7	IV	IV

Predictions made by our classifier based on motifs found by a variety of conventional motif finding algorithms in the context of genome-wide location analysis. We considered only reported motifs whose TFs were known and fell into one of the six classes, but did not belong to the training set.

computational scan of bound genomic sequences with a PSSM learned from those same sequences. We provided the resultant feature vectors to our classifier to see if it could correctly determine the class in each case.

Table 4 shows the results of the test. Only 3 of the 14 yeast factors—Fhl1, Phd1 and Sok2—are misclassified. Subsequent analysis suggests that the errors associated with Phd1 and Sok2, both bHLH proteins, may be related to an error in the motif identified by Harbison *et al*. in their paper. The reported motifs for these proteins have a high similarity to the motif of Sut1, a type of zinc-coordinating TF (not in the six classes we have considered). The algorithms used by Harbison *et al*. may learn this motif for Phd1 and Sok2 because Sut1 binds to many of the genomic sequences to which these factors bind. The Sut1 motif appears to be strong enough to dominate in each TF's sequence set. Consequently, it is not clear if the reported motifs of these two TFs are correct. In summary, we correctly predict the domains of 11 TFs out of either 14 or 12, depending on the correctness of the two bHLH motifs. Either way, our classifier seems to hold promise as a tool for supplementing conventional motif finding algorithms, predicting the class of TF that binds the sequences from which a motif is constructed.

5 DISCUSSION

We present a novel classifier that classifies TFs into classes defined by their structural DNA-binding domains using only features of the sequences of their binding sites. We demonstrate that the sequence features of TF binding sites contain significant predictive information to reveal the structural mechanisms of TF binding.

Taking a closer look at the features selected by the classifier provides insight into the various sequence features favored by each class. The classifier we learn from the full training set selects a total of 323 features. Out of these, nearly a third (107) are dominant predictors of Class I, more than any other class. Class I contains Cys_2His_2 zinc-coordinating TFs that contact DNA with a series of zinc fingers, each of which are believed to bind to three or four nucleotides somewhat independently. Given that we consider all possible triplets and quadruplets in our feature vector, one would expect this class to be well distinguished from the others. However, on the contrary, almost 23% are misclassified. Also, 40% of all the errors in the other five classes are due to TFs in those classes being misclassified as Class I. Indeed, if we remove Class I from our training set, and perform LOOCV on the 490 TFs belonging to the other five classes, we achieve an overall error rate of only 9%, less than three-quarters of what we achieve with all six classes. In addition, the number of selected features decreases to 241, about three-quarters of the number with all six classes. The fact that Class I is relatively poorly characterized compared with other classes can be explained by looking at how Cys_2His_2 TFs bind to DNA. Individual zinc fingers have poor sequence specificity (Pabo and Sauer, 1992), which explains why multiple fingers are needed for the TF to bind to DNA with greater affinity and specificity. This further implies that a large number of distinct binding sites are possible due to the combinations of the recognition sites for each finger. Because of the several variable triplets and quadruplets, it appears that the classifier selects many substrings, but is not able to define the correct weights to be assigned to the feature vector for Class I. Thus, although the binding mechanism of all the TFs in this class is the same, it is difficult to distinguish this class from others based solely on the sequence features considered by the classifier. It is of course possible that expanding the feature set may result in better classification, from the perspective of accuracy or sparsity.

Only two of the palindrome-related features are selected consistently (586 times) across all the 587 classifiers. Both are dominant predictors of Class II, the other class of zinc-coordinating proteins. Most members of this class are known to form dimers, which explains their preference for palindromic sites. This is the only class that especially favors gapped palindromes. If we look at the binding mechanism of a Cys_4 TF with the DNA, we notice that many Cys_4 proteins form dimers after binding to DNA (Lefstin and Yamamoto, 1998), which may explain the preference of these TFs for gapped palindromes. In contrast to Class I, this class has the smallest number of features that are dominant predictors of the class; yet, this small number of features suffices to achieve high accuracy in predicting this class of TFs.

Not surprisingly, the special features we included in the feature vector (see Table 1) were consistently selected as dominant predictors of the corresponding classes; in the case of Cys_2His_2, which had three special features, only `[GC]..[GC]..[GC]` was selected.

Our results have larger implications both for how TF binding sites should be modeled and for how algorithms should be developed for finding them. Regarding the former, because of the importance of gapped palindromes in the binding sites of certain classes of TFs, having a fixed length model is probably not the best way to represent motifs. As just one specific but not uncommon example, consider the mouse transcription factor JunD (TRANSFACid: T00437), belonging to the bZip family. According to TRANSFAC, it has eight known binding sites, four of which are the following:

(1) `ATGACTCAT`

(2) `ATGACGTCAT`

(3) TGACATCA

(4) TGACTAA

These four binding sites have different lengths, and even if the sites were padded at the ends to ensure a uniform length, the consensus triplets TGA and TCA would be separated by a variable number of bases; this shows that no PSSM model can effectively capture their important structure. The same concern applies to more flexible probabilistic representations that were introduced to improve upon PSSMs but continue to assume a fixed length binding site (e.g. weight array matrices, trees, Bayesian networks). On the other hand, non-probabilistic regular expression models might have difficulty in sensitively identifying the critical gapped palindromic site TGA.*TCA because of the sequence variability present in the fourth binding site. JunD is, however, correctly recognized as a bZip factor by our classifier. We are working to develop a probabilistic representation of TF binding sites that will accommodate length variability and maintain local alignments, while at the same time, take into account the structural class of a TF. Hidden Markov models and probabilistic regular or context-free grammars seem promising.

We have also demonstrated that our classifier can be used to predict the class of TF binding to putative motifs discovered with conventional motif finding algorithms. But since various sequence features are selected uniquely for different classes of TFs, this kind of information might also be useful in improving the motif finding algorithms themselves. The only papers of which we are aware that have taken a step in this direction are those of Xing and Karp (2004) and Sandelin and Wasserman (2004). Xing and Karp (2004) use structure-specific features to generate a prior over PSSM models while Sandelin and Wasserman (2004) use pseudo-counts calculated from their familial binding profiles as a prior over PSSM models. Our results suggest that features within the sequences of promoters of co-regulated genes can be used to produce a prior over positions within the promoter sequences. Indeed, we have already produced a new algorithm for *de novo* motif finding that uses a classifier related to the one reported here to construct an informative prior over sequence positions, and then uses collapsed Gibbs sampling (Liu, 1994) to sample from the posterior (Narlikar *et al*., 2005).

In closing, we note also that the demonstrated connections between DNA sequence features and TF structural classes should also be useful in engineering proteins that bind specifically to particular DNA sequences. For example, our classifier might predict that a certain target DNA sequence will be poorly recognized by some classes of TFs, but well recognized by others; in such a setting, using our classifier to learn in advance the class of TF that ought to be engineered might save a considerable amount of time.

ACKNOWLEDGEMENTS

The authors would like to thank Balaji Krishnapuram who wrote the original code for the sparse multinomial logistic regression (SMLR) algorithm. A new version of the code, written by Jason Bosko, Matthew Edwards, and Alexander Hartemink, is available for non-commercial use from http://www.cs.duke.edu/~amink/. This research was supported in part by an NSF CAREER award and the Sloan Foundation Fellowship to A.J.H.

Conflict of Interest: none declared.

REFERENCES

Agarwal,P. and Bafna,V. (1998) Detecting non-adjoining correlations within signals in DNA. In *Proceedings of RECOMB '98*, 2–8.

Aitola,M. (2002) Developmental expression of transcription factors. *Academic Dissertation*. Medical School of University Tampere, Finland.

Alberts,B., Johnson,A., Lewis,J., Raff,M., Roberts,K. and Walter,P. (2002) *Molecular Biology of the Cell*. 4th edn, Garland Science, New York.

Atchley,W. and Fitch,W. (1997) A natural classification of the basic helix loop helix class of transcription factors. *Proc. Natl Acad. Sci. USA*, **94**, 5172–5176.

Bailey,T. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *ISMB '94*, AAAI Press, Menlo Park, CA, pp. 28–36.

Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003) Modeling dependencies in protein-DNA binding sites. In *Proceedings of RECOMB '03*, 28–37.

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.

Derreumaux,S. and Fermandjian,S. (2000) Bending and adaptability to proteins of the cAMP DNA-responsive element: molecular dynamics contrasted with NMR. *Biophys. J*., **79**, 656–669.

Harbison,C. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hertz,G. and Stormo,G. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Kellis,M. *et al*. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **432**, 241–254.

Krishnapuram,B. *et al*. (2005) Learning sparse Bayesian classifiers: multi-class formulation, fast algorithms, and generalization bounds. *IEEE Trans. Pattern Anal. Machine Intell*., **27**, 957–968.

Lefstin,J. and Yamamoto,K. (1998) Allosteric effects of DNA on transcriptional regulators. *Nature*, **392**, 885–888.

Liu,J. (1994) The collapsed Gibbs sampler with applications to a gene regulation problem. *J. Amer. Stat. Assoc*., **89**, 958–966.

Liu,X. *et al*. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin immunoprecipitation microarray experiments. *Nat. Biotechnol*., **20**, 835–839.

Mulder,N. *et al*. (2003) The InterPro Database: 2003 brings increased coverage and new features. *Nucleic Acids Res*., **31**, 315–318.

Narlikar,L., Uwe,O., and Hartemink,A. (2005) Informative priors improve motif discovery, (in preparation).

Pabo,C. and Sauer,R. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem*., **61**, 1053–1095.

Roth,F. *et al*. (1998) Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol*., **16**, 939–945.

Sandelin,A. *et al*. (2004) JASPAR: an open access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*., **32** (Datebase issue), D91–D94.

Sandelin,A. and Wasserman,W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol*., **338**, 207–215.

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

Williams,P. (1995) Bayesian regularization and pruning using a Laplace prior. *Neural Comput*., **7**, 117–143.

Wingender,E. *et al*. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*., **29**, 281–283.

Wolfe,S. *et al*. (2000) DNA recognition by Cys_2His_2 zinc finger proteins. *Annu. Rev. Biomol. Struct*., **3**, 183–212.

Xing,E. and Karp,R. (2004) MotifPrototyper: a Bayesian profile model for motif families. *Proc. Natl Acad. Sci. USA*, **101**, 10523–10528.

Zhang,M. and Marr,T. (1993) A weight array method for splicing signal analysis. *Comput. Appl. Biosci*., **9**, 499–509.

Zilliacus,J. *et al*. (1995) Structural determinants of DNA-binding specificity by steroid receptors. *Mol. Endocrinol*., **9**, 389–400.