Motif Discovery as a Multiple-Instance Problem

Ya Zhang
Department of EECS
University of Kansas
yazhang@ittc.ku.edu

Yixin Chen
Department of CIS
The University of Mississippi
ychen@cs.olemiss.edu

Xiang Ji
Yahoo! Inc.
xiangji@yahoo-inc.com

Abstract

Motif discovery from biosequences, a challenging task both experimentally and computationally, has been a topic of immense study in recent years. In this paper, we formulate the motif discovery problem as a multiple-instance problem and employ a multiple-instance learning method, the MILES method, to identify motif from biological sequences. Each sequence is mapped into a feature space defined by instances in training sequences with a novel instance-bag similarity measure. We employ 1-norm SVM to select important features and construct classifiers simultaneously. These high-ranked features correspond to discovered motifs. We apply this method to discover transcriptional factor binding sites in promoters, a typical motif finding problem in biology, and show that the method is at least comparable to existing methods.

1 Introduction

With the increasing volume of biological sequences available, an important bioinformatics problem is to find regularities among the sequences as motifs. In biological sequence analysis, a motif is a short consensus patterns among a set of biological sequences and it represents a common feature or inherent pattern in the sequences. For proteins, motifs are generally closely related to their functions and structures. DNA motifs are often present at the non-coding region of genome and serve as signals to determine interactions between DNA, RNA transcripts, and the cellular machinery. Discovering these motifs plays an important role in understanding how cell functions. In its simplest form, the motif discovery problem can be generally formulated as follows.

> Given a set of sequences, each of which are known to be embedded at least one instance of a motif of length l with up to d mutations, recover the motif.

The development of automated methods for motif finding in biosequence is a challenging task for the following reasons. First, because the motif instances are subject to various kinds of mutations such as substitution, deletion and insertion, the instances could significantly different from each other although each instance still closely matches the consensus pattern. For example, suppose the motif consensus is $CTTCCT$. Two motif instances, $ACTCCT$ and $CTTTCT$, each with only one or two mutations from the motif consensus, show very subtle similarity. Secondly, some other false signal similar to the motif consensus may randomly occur in the sequences and obscures the true motif's signal.

In the past several years, many motif discovery algorithms have been proposed based on greedy algorithms [9], Expectation Maximization(EM) algorithms [1, 4], Gibbs Sampling [12, 13, 19], evolutionary computation [14], and many other algorithms. Depending on the way of defining the motif search space, they generally fall into two categories: pattern-driven approaches [17, 16] and sample-driven approaches [4, 12, 13, 19, 18]. A hybrid of pattern-driven and sample-driven approaches, the MULTIPROFILER algorithm, was also proposed [10], where a neighborhood of each segment in the DNA sequences is used as a possible motif. Sinha [16] proposed Dmotif algorithm to address the problem from a feature selection perspective. Each candidate motif is viewed as a feature. Classifiers are built on each of the features to discriminate positive sequences from random sequences. Candidate motifs with the smallest classification errors are reported as likely patterns.

In this paper, we formulate the motif discovery problem in bioinformatics as a multiple-instance problem. A *multiple-instance problem* involves ambiguous training examples: a single example is represented by a set of instances, some of which may be responsible for the observed classification of the example; yet, the training label is only attached to the example instead of the instances. In the case of motif discovery, sequences embedded with a certain motif of length are positive examples and random sequences are used as negative examples. Suppose the length of the motif is l. Each sequence is represented by a collection of instances – all the unique overlapping l-mers in

the sequence. The motif discovery problem is similar to the multiple-instance problem in the sense that each positive example is known to have one or more positive motif instances embedded. However, different from a typical multiple-instance problem, in motif discovery, a true motif may also randomly appear in a negative sequences. Hence algorithms which implicitly assume that instances in negative examples are all negative may not work in this case. The MILES (Multiple-Instance Learning via Embedded instance Selection) method [5] which converts the multiple-instance learning problem to a standard supervised learning problem does not impose the assumption relating instance labels to example labels and hence is suitable for the motif discovery problem. We apply this method to discover transcriptional factor binding sites in promoters, a typical motif finding problem in biology, and show that the method is at least comparable to existing methods.

2 Review of Multiple-Instance Learning

Multiple-instance learning differs from supervised learning in that labels are only assigned to collections of instances (bags) rather than individual instances. Generally, a bag is labeled positive if and only if at least one instance in that bag is positive. Otherwise it is labeled negative. No label is given to individual instances. The goal of multiple-instance learning is to discover instances that are responsible for positive labeling of the bags using the labeled bags in training data and hence classify new bags.

One of the earliest algorithms for multiple-instance learning is axis-parallel rectangle (APR) method proposed for drug activity prediction [7]. The idea of the APR method was extended to a general framework based on *diverse density* (DD) [15], which measures a co-occurrence of similar instances from different positive bags. Zhang and Goldman [21] combined the idea of expectation-maximization (EM) with diverse density, and developed the EM-DD algorithm to search for the most likely concept. Ensembles of multi-instance learners were also proposed [22], which achieved competitive test results in drug activity prediction. Multiple-instance problems have also been addressed with standard supervised learning techniques, including decision trees [25], Support Vector Machines (SVMs) with a kernel for multiple-instance data [8], logistic regression and boosting approaches [20], and SVMs with DD function [6].

Many of the above multiple-instance formulations explicitly or implicitly encode the assumption that a bag is positive if and only if at least one of its instances is positive. The assumption is valid for the earliest studied multiple instance problems such as drug activity prediction. However, for applications such as motif discovery, a negative bag (e.g., a random sequence without the motif embedded) may also contain instances that are similar to the motif purely by random. The MILES (Multiple-Instance Learning via Embedded instance Selection) method [5], which converts the multiple-instance learning problem to a standard supervised learning problem that does not impose the assumption relating instance labels to bag labels, is well suited for the motif discovery task. This method maps each bag into a feature space defined by the instances in the training bags via an instance-bag similarity measure and apply 1-norm SVM to select important features as well as construct classifiers simultaneously.

3 MILES for Motif Discovery

MILES [5] extends ideas from the diverse density framework [15, 6] and the wrapper model in feature selection [11]. It identifies instances that are relevant to the observed classification by embedding bags into an instance-based feature space and selecting most important features. Based on an instance-bag similarity measure, a given bag is embedded into a feature space where each dimension represents the bag's similarities to a particular instance in the training set. The embedding produces a possibly high dimensional space when the number of instances is large. In addition, many features may be redundant or irrelevant because some of the instances might not be responsible for the observed classification of the bags, or might be similar to each other. It is hence essential and indispensable to select a subset of mapped features that is most relevant to the classification problem of interest. 1-norm SVM [3, 23] is applied to construct classifiers and select important features simultaneously. Since each feature is defined by an instance, feature selection is essentially instance selection. The selected instances define the motifs uncovered.

We denote positive sequences (bags) as s_i^+ and the j-th l-mers (instance) in that sequences as s_{ij}^+. The sequence s_i^+ consists of n_i^+ instances s_{ij}^+, $j = 1, \cdots, n_i^+$. Similarly, s_i^-, s_{ij}^-, and n_i^- are defined for random sequences (negative bags). When the label on the sequences does not matter, it will be referred to as s_i with l-mers as s_{ij}. All l-mers belong to feature space $\mathbb{X}$. The number of positive (negative) sequences is denoted as ℓ^+ (ℓ^-). For the purpose of motif discovery, we are interested in finding positive instances. Hence, instances that only occur in negative bags are ignored. For the sake of convenience, we line up all instances in positive bags together and reindex them as $\mathbf{x}^k$, $k = 1, \cdots, n$, where $n = \sum_{i=1}^{\ell^+} n_i^+$.

Next, we introduce a novel instance-bag similarity measure designed specifically for motif discovery. We define the alignment score between the instance c and the sequence s_i as the maximum alignment score between the instance c and all l-mers in the sequence s_i.

$$A(c, s_i) = \max_{s_{ij} \in s_i} A(c, s_{ij}) \,, \tag{1}$$

where $A(c, s_{ij})$ is the ratio of matched positions when the two l-mers. The similarity between an instance and a se-

quence is proportion to their alignment score. This definition has a winner-takes-all flavor, in that, as long as the sequence s_i contains the instance, the winner will be the segment corresponding to the instance, and the similarity will be large. Intuitively, given an instance c, the probability that the sequence s_i are embedded with instance c is high if we can find a close match for c at the sequence.

Because the biological sequences are assumed to be generated from a random background distribution, it is reasonable to assume that an instance would more likely to be a positive instance if the probability that it is generated from the background model is low. Therefore the similarity between an instance and a sequence is inverse proportion to the probability that the instance c is generated by the background model ($P(c)$). The length of the instance l accounts for $P(c)$ (longer instances generally have lower probability). To eliminate this factor, we normalize the probability $P(c)$ with the length of the instance l. The similarity between an instance c_j and a sequence s_i is therefore formulated as

$$m_{ij} = e^{\alpha A(s_i, c_j) - \log P(c_j)/l}. \quad (2)$$

In this way, we map each sequence into a feature space defined by the instances in the training examples via an instance-bag similarity measure. This feature mapping often provides a large number of redundant or irrelevant features. We apply 1-norm SVM to select important instances (features) as well as construct classifiers of the examples simultaneously.

Next we present a brief review the 1-norm SVM formulation. The class label of the sequences are denoted by y, which takes values of $+1$ and -1 ($+1$ for those sequences with motif embedded and -1 for those without). We consider the classification problem of finding a linear classifier $y = \text{sign}\left(\mathbf{w}^T\mathbf{m} + b\right)$ in the feature space $\mathbb{F}_\mathcal{C}$ to distinguish between positive examples and negative examples where $\mathbf{w}$ and b are model parameters, $\mathbf{m} \in \mathbb{F}_\mathcal{C}$ corresponds to a bag. The SVM approach constructs classifiers based on hyperplanes by minimizing a regularized training error $\lambda P[\cdot] + error$ where $P[\cdot]$ is a regularizer, λ is called the regularization parameter, and $error$ is commonly defined as a total of the loss that each bag introduces through a hinge loss function $\xi = \max\{1 - y(\mathbf{w}^T\mathbf{m} + b), 0\}$. When an optimal solution $\mathbf{w}$ is obtained, the magnitude of its component w_k indicates the significance of the effect of the k-th feature in $\mathbb{F}_\mathcal{C}$ on the classifier. Those features corresponding to a non-zero w_k are selected and used in the classifier.

The regularizer in standard SVMs is the squared 2-norm of the weight vector $\|\mathbf{w}\|$, which formulates SVMs as quadratic programs (QP). Solving QPs is typically computationally expensive. Alternatively, SVMs are formulated as Linear programs (LPs) [2, 23] by regularizing with a sparse-favoring norm, e.g., the 1-norm of $\mathbf{w}$ ($\|\mathbf{w}\|_1 = \sum_k |w_k|$). Thus 1-norm SVM is also referred to as sparse SVM and has been applied to other practical problems such as drug discovery [3]. By rewriting $\mathbf{w}$ as $\mathbf{w} = \mathbf{u} - \mathbf{v}$, the LP for 1-norm SVM can be formulated as:

$$\begin{aligned}
\min_{\mathbf{u},\mathbf{v},b,\boldsymbol{\xi},\boldsymbol{\eta}} \quad & \lambda \sum_{k=1}^{n} (u_k + v_k) + \mu \sum_{i=1}^{\ell^+} \xi_i + (1-\mu) \sum_{j=1}^{\ell^-} \eta_j \\
\text{s.t.} \quad & \left[(\mathbf{u} - \mathbf{v})^T \mathbf{m}_i^+ + b\right] + \xi_i \geq 1, i = 1, \cdots, \ell^+, \\
& -\left[(\mathbf{u} - \mathbf{v})^T \mathbf{m}_j^- + b\right] + \eta_j \geq 1, j = 1, \cdots, \ell^-, \\
& u_k, v_k \geq 0, k = 1, \cdots, n, \\
& \xi_i, \eta_j \geq 0, i = 1, \cdots, \ell^+, j = 1, \cdots, \ell^-. \quad (3)
\end{aligned}$$

where $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ are hinge losses. Let $\mathbf{w}^* = \mathbf{u}^* - \mathbf{v}^*$ and b^* be the optimal solution of (3). The magnitude of w_k^* determines the influence of the k-th feature on the classifier. The set of selected features is given as $\{\mathbf{x}^k : k \in \mathcal{I}\}$, where $\mathcal{I} = \{k : |w_k^*| > 0\}$ is the index set for nonzero entries in $\mathbf{w}^*$.

4 Experiments and Results

In this section, we present the experiments of applying MILES to a specific motif discovery problem – discovering transcriptional factor binding sites at Yeast promoter regions. The binding sites of transcriptional factors usually display a motif pattern. The Promoter Database of *Saccharomyces Cerevisiae* (SCPD)[1] catalogs more than 100 transcriptional factors. For each transcriptional factor, it provides information about genes under its regulation, experimentally-mapped binding sites, as well as a consensus binding site. To test the performance of our algorithm, test sets are built from SCPD for transcriptional factors occurring in no less than 3 promoters and having a consensus binding site. Totally 22 transcriptional factors are selected, and each data set contains the promoter sequences of the corresponding genes and consensus binding site. The length of the promoter sequences is 600 bps. A 3^{rd}-order Markov chain model is built for background sequences with all yeast promoter sequences that are publicly available[2].

4.1 Experiments

We apply the MILES algorithm to discovering the experimentally-mapped transcriptional factor binding sites in the 22 sets of promoter sequences. For each transcriptional factor, we are given a set of DNA sequences which the transcriptional factor is known to bind to. These sequences serve as the positive examples (bags) for multiple-instance learning. We randomly generate an equal number of sequences of the same length using the above background model and use them as the negative examples (bags) for the learning task. Two key parameters are required by the algorithm: the length of motif instance l and the length of spaces k. A sliding window of size l is used to extract instances from positive sequences. We also allow spacer in

[1] *http://cgsigma.cshl.org/jian/index.html*

[2] The sequences retrieval tool provided by SCPD is used to retrieve yeast promoters to estimate the parameters for the background model.

the motif. According to previous findings [16, 24], if a motif contains spacer, the spacer usually presents at the middle of the motif with length from 1 to 11 base pairs. Suppose k is the length of spacer allowed. We mask k consecutive letters at the center of each instance.

4.2 Results

The experimental results are summarized in Table 1. Column 1 lists the names of the transcriptional factors. Column 2 shows the biologically-mapped binding site consensus of the corresponding transcriptional factors. The first close match to the consensus binding site and its rank, reported by the MILES algorithm, are presented in columns 3 and 4. The first close match reported by Dmotif algorithm and its rank are given in columns 5 and 6. It can be seen that in 11 out of 22 categories of promoters, the known consensus closely matches the top ranking motifs reported by MILES algorithm.

We here compare the prediction results of MILES algorithm with the published results of Sinha's Dmotif algorithm [16]. We compute the percentage of cases in which the first-ranked motif reported by the algorithms closely matches the known consensus. For our algorithm, this value is 50%, while the value for Dmotif algorithm is 45%. Thus, in terms of the percentage of first-ranked true motifs, our algorithm performs comparably to the Dmotif algorithm. In nine out of 22 cases, our algorithm assigns the experimentally determined binding sites a higher rank than the Dmotif algorithm, while in five out of 22 cases, Dmotif algorithm ranks them higher.

In addition, our algorithm also successfully recover the binding sites of the regulons TBP and $UASPHR$, which Dmotif algorithm failed to identify. The highest-ranked motif pattern $TATAAA$ closely matches the experimentally identified binding sites of TBP. $CTTCCT$, which exactly matches the binding site $CTTCCT$ of $UASPHR$, is reported at rank 15. Both our proposed algorithm and Dmotif algorithm are unable to discover the annotated motif consensus of regulon SFF.

5 Discussion and Conclusions

Discovering motifs from biological sequences is a very difficult task. In this paper, we have formulated the motif discovery problem as a multiple instance problem and applied a multiple-instance learning algorithm, the MILES algorithm, to finding motifs. Using a novel an instance-bag similarity measure, we map each sequence into a feature space defined by the instances in the training sequences. We employ 1-norm SVM to select important features as well as construct classifiers simultaneously. Our experimental results have shown that this algorithm is capable of ranking the true motifs as top matches in 50% of the cases as compared with 45% of the cases with Dmotif algorithm.

The motif length and the length of spaces are given as parameters for our motif discovery algorithm. Although the length of the motif is in general 6 to 8, the possible range of the length of spaces could be large. One extension of this work is to automatically find the two parameters.

In this paper, we limit the motif search space to the patterns that actually appear in the positive sample sequences, as sample-driven algorithms do. However, in the cases that variations are allowed in the sequences or in the cases that most motif instances are differently mutated and hence significantly from each other, sample-driven approach may not work well. One way to avoid the problem is to construct a feature space using a neighborhood of each segment in the DNA sequences as a possible motif instance.

References

[1] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21(1/2):51–80, 1995.

[2] K. P. Bennett. *Advances in Kernel Methods – Support Vector Machines*, chapter Combining Support Vector and Mathematical Programming Methods for Classification, pages 307–326. 1999.

[3] J. Bi, K. P. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3:1229–1243, 2003.

[4] J. Buhler and M. Tompa. Finding motifs using random projections. In *Proc. RECOMB*, pages 69–75, Montreal, Canada, 2001.

[5] Y. Chen, J. Bi, and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, to appear*, 2006.

[6] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939, 2004.

[7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[8] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *Proc. 19th Int'l Conf. on Machine Learning*, pages 179–186, 2002.

[9] G. Z. Hertz and G. D. Stormo. Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15:563–577, 1999.

[10] U. Keich and P. Pevzner. Finding motifs in the twilight zone. In *Proc. RECOMB*, pages 195–204, Washington, D.C., USA, 2002.

[11] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[12] J. S. Liu, A. F. Neuwald, and C. E. Lawrence. Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *J. Am. Stat. Assoc*, 90(432):1156–1170, 1995.

Table 1. Results of predicting transcriptional factor binding sites in yeast promoter regions. The results of Dmotif algorithm [16] are included here for ease of comparison. The consensus binding sites from SCPD and the results of Dmotif algorithm are built upon International Union Of Pure And Applied Chemistry (IUPAC) degenerate symbols ($\{$A,C,G,T,R,S,W,M,Y,K,N$\}$) that are restricted expressions over $\{$A, C, G, T$\}$. * The reported motif overlaps with the binding sites cataloged in SCPD. Thus, it is considered a close match.

REGULON	BINDING SITE	MILES		DMOTIF	
		MOTIF FOUND	RANK	MOTIF FOUND	RANK
ABF1	TCRNNNNNNACG	CACNNNNNNCGT	4	TCANNNNNNAMG	2
CPF1	TCACGTG	CACGTGG	1	CACGTG	1
CSRE	YCGGAYRRAWGG	ACGGATAG	7	CGGATGRA	8
SCB	CNCGAAA	GTCACGA	1	TCGCGAA	2
GAL4	CGGNNNNNNNNNNCCG	CGGNNNNNNNNNNCCG	1	CGGNNNNNNNNNNCCG	1
GCR1	CWTCC	CCTTC	7	CTTCC	13
HAP1*	CGGNNNTANCGG	GGGNNNNNCGG	2	GGANNNNNCGG	1
HSE	TTCNNGAA TTCNNNGAA GAANNTCC GAANNNTCC	TTCTAGAA	1	TTMTAGAA	6
MCB	WCGCGW	CGCGTG	2	ACGCGT	1
MCM1*	CCNNNWWRGG	CCTAATTGGG	4	TTTCCTAA	1
MATa2	CRTGTWWWW	CATGTAAT	1	CATGTMA	2
M1G1	CCCCRNNWWWWW	CCCCAG	1	MCCCCAG	1
PHO4	CACGTK	CACGTG	1	CACGTG	1
PDR3	TCCGYGGA	TCCGCGGA	1	TCCGYGGA	2
REB1	YYACCCG	TACCCGC	1	YTACCCG	1
ROX1	YYNATTGTTY	GCCTATTGTT	1	CCTATTG	7
RAP1	RMACCCA	GAACCCA	5	ACCCAGW	1
CAR1	AGCCGCSA	TAGCCGC	2	TAGCCGCS	2
SFF	GTMAACAA	NOT FOUND	-	NOT FOUND	-
STE12	ATGAAA	ATGAAAC	1	ATGNAAC	1
TBP	TATAWAW	TATAAA	3	NOT FOUND	-
UASPHR	CTTCCT	CTTCCT	15	NOT FOUND	-

[13] X. Liu, D. L. Brutlag, and J. S. Liu. Bioprospector: Discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. In *Proc. of Pac. Symp. Biocomput*, pages 127–38, 2001.

[14] M. A. Lones and A. M. Tyrrell. The evolutionary computation approach to motif discovery in biological sequences. In *Proc. Workshop on Biological Applications of Genetic and Evolutionary Computation (BioGEC), GECCO2005*, 2005.

[15] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. *Advances in Neural Information Processing Systems 10*, 10:570–576, 1998.

[16] S. Sinha. Discriminative motifs. In *Proc. of RECOMB*, pages 291–298, Washington, D.C., USA, 2002.

[17] S. Sinha and M. Tompa. A statistical method for finding transcription factor binding sites. In *Proc. of ISMB*, pages 344–354, 2000.

[18] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. Moor, P. Rouze, and Y. Moreau. A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. *Bioinformatics*, 17(12):1113–1122, 2001.

[19] G. Thijs, K. Marchal, and Y. Moreau. A gibbs sampling method to detect over-represented motifs in the upstream regions of co-expressed genes. In *Proc. of RECOMB*, pages 305–312, Montreal, Canada, 2001.

[20] X. Xu and E. Frank. Logistic regression and boosting for labeled bags of instances. In *Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, pages 272–281, 2004.

[21] Q. Zhang and S. A. Goldman. EM-DD: An improved multiple-instance learning technique. *Advances in Neural Information Processing Systems 14*, 14:1073–1080, 2002.

[22] Z.-H. Zhou and M.-L. Zhang. Ensembles of multi-instance learners. *Lecture Notes in Artificial Intelligence*, 2837:492–502, 2003.

[23] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. *Advances in Neural Information Processing Systems 16*, 2004.

[24] J. Zhu and M. Q. Zhang. Scpd: a promoter database of the yeast *saccharomyces cerevisiae*. *Bioinformatics*, 15(7/8):607–611, 1999.

[25] J.-D. Zucker and Y. Chevaleyre. Solving multiple-instance and multiple-part learning problems with decision trees and rule sets, application to the mutagenesis problem. *Lecture Notes in Artificial Intelligence*, 2056:204–214, 2001.