Identifying Conserved Discriminative Motifs

Jyotsna Kasturi[1,2], Raj Acharya[2], and Ross Hardison[3,4]

[1] Non-Clinical Biostatistics, Johnson & Johnson Pharmaceutical Research & Development, New Jersey
[2] Department of Computer Science and Engineering
[3] Center for Comparative Genomics and Bioinformatics, Huck Institutes of Life Sciences
[4] Department of Biochemistry and Molecular Biology; Pennsylvania State University
jkasturi@its.jnj.com, acharya@cse.psu.edu, rch8@psu.edu

Abstract. The identification of regulatory motifs underlying gene expression is a challenging problem, particularly in eukaryotes. An algorithm to identify statistically significant discriminative motifs that distinguish between gene expression clusters is presented. The predictive power of the identified motifs is assessed with a supervised Naïve Bayes classifier. An information-theoretic feature selection criterion helps find the most informative motifs. Results on benchmark and real data demonstrate that our algorithm accurately identifies discriminative motifs. We show that the integration of comparative genomics information into the motif finding process significantly improves the discovery of discriminative motifs and overall classification accuracy.

Keywords: Discriminative motifs, regulatory elements, comparative genomics, classification, Naïve Bayes, mutual information.

1 Introduction

One of the challenges of post-genomic molecular biology is to understand gene regulation and the complex mechanisms underlying gene expression. In protein-coding genes, gene regulation occurs by altering the rate of transcription of the gene thereby changing its expression levels. Transcriptional regulatory proteins exert control on the expression of genes in a cell by binding to specific DNA regulatory elements, in close proximity to the transcription start site of a gene. Computational methods have been quite successful in identifying patterns in DNA sequences as putative transcription factor binding sites, especially in bacteria and yeast. These methods typically make use of available gene expression measurements either to guide the search for motifs within DNA sequences [2], [3], [5], [13], or in a combined fashion by correlating gene expression data with the sequences [6], [7].

Eukaryotic binding site prediction remains a complex problem and a challenging one for several reasons. The first is that transcription factors often bind to regions of the DNA several kilobases from the gene's transcription start site. These sites may even be present within the introns or downstream of the gene. The factor MEF2 in humans, and GATA1 in mouse and human are examples. Another critical issue is that motif length is unknown and often highly variable between sites although they often

M. Chetty, A. Ngom, and S. Ahmad (Eds.): PRIB 2008, LNBI 5265, pp. 334–348, 2008.

have a common core site of smaller length (such as WGATAR). In addition, transcription factors may act cooperatively with other factors to control regulation. Since the length of sequences can be very long, computational identification of putative binding sites can lead to many false positives being detected. Functional regulatory elements are often conserved over evolution and true binding sites will often be present in highly conserved regions of the multi-species sequence alignments. Some motif finding algorithms have used this approach to identify putative binding sites [1], [9].

Motifs are represented using either position weight matrices (PWM) or string-based models, and identified as over-represented patterns common to a given set of sequences relative to a set of background sequences. Methods that model motifs with a probability matrix model [11], [12], [19] are capable of capturing variability between binding sites effectively, but use maximum likelihood and local search techniques to estimate parameters which are extremely dependent on starting conditions, number of iterations and stability of the model to converge to a local maximum/minimum. String-based motif finding algorithms [16], [17] on the other hand are based on basic counting schemes and have the advantage of being deterministic, in the sense that they produce identical results for every run with a fixed set of parameters. However, the motifs identified will vary depending on the choice and estimation method of the background distribution.

Discriminative methods find motifs that distinguish between two sets of input sequences or gene clusters, thereby avoiding the need for a background distribution. A probabilistic logistic regression model was proposed to identify discriminative motifs between two sets of genes, one expected to contain a *cis*-regulatory module (CRM) while the other did not [14]. The algorithm performs well but requires the estimation of an extremely large number of variables, with complexity exponential in the length of the sequences. Also, prior knowledge of CRMs is generally unavailable. DMotifs [17] uses a string-based approach to look for well-distributed motifs over individual promoters, identifying discriminative motifs between a set of positive and negative (background) sequences. The algorithm does not scale to large number of sequences. The DME algorithm [18] also uses an enumerative approach to perform an exhaustive search within a set of candidate matrices; refining the highest scoring matrices and erasing discovered motifs from the data iteratively. These methods cannot handle multiple clusters.

We present a supervised learning approach to identify conserved discriminative motifs between multiple cohorts of genes. An enumerative method is used to model each gene individually based on the distribution of words in the sequence and distinguishing between words with different counts. Words (putative motifs) that do not contribute to the discrimination as measured by a feature selection method are iteratively dropped, resulting in a drastic reduction of the search space and number of candidate motifs. We show that the discriminative power of the identified motifs is increased by 15%-20% when comparative genomics information is incorporated into the algorithm. The design and implementation of the algorithm is specifically targeted to handle very large amounts of data.

2 Methods

Let K be the number of clusters into which the N input genes have been grouped based on gene expression or function similarity. Let S be the set of DNA sequences corresponding to the genes consisting of upstream, intron and downstream regions for each gene. The objective of the discriminative algorithm is to identify a small set of motifs or word patterns that can discriminate between these input gene clusters (Fig. 1). For example, given a set of genes divided into an up-regulated cluster and a down-regulated cluster, we are interested in identifying those motifs that discriminate between the two clusters thereby gaining a better understanding of their individual regulatory mechanisms. Once a relatively small set of putative motifs has been identified using computational means, they may be validated through appropriate in-vivo and in-vitro assays.

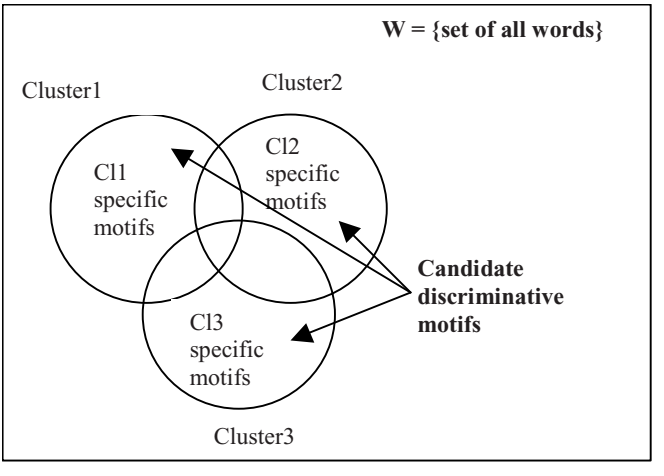

Fig. 1. Venn diagram showing candidate motifs to discriminate between 3 clusters

2.1 The Mathematical Motif Model

Let gene z correspond to a set of sequences (upstream, intron and downstream) denoted by s^z; $z = 1, 2, \ldots, N$. We use a string-based motif model that is capable of distinguishing between distinct words and their frequencies of occurrence. For a specified motif length l, the sequences are scanned using a sliding window of the corresponding size to obtain a list of all the word patterns. In defining the motif model, we distinguish between the terms *token* and *word*. A *token* is a DNA string pattern of specified length l. Let $V=\{t_1, t_2, \ldots, t_{|V|}\}$ denote the token vocabulary for the dataset. Given a sequence, the token t_i may appear with frequency count $n(t_i) >= 0$. Let the model parameter R (>= 1) denote the maximum frequency count considered by the model. The modified count x_i for token t_i is then given by

$$x_i = \begin{cases} n(t_i) & \text{if } n(t_i) \in \{0,1,...,R\} \\ R & \text{if } n(t_i) > R \end{cases} \quad (1)$$

A *word* (potential motif) is then defined as the tuple(t, x) consisting of a token and its modified frequency, where $t \in V$ and $x \in X = \{0,1,2,\ldots,R\}$. Let $W \subset V \times X$ denote the set of all words (vocabulary of words) present in the dataset. Any given sequence s^z may then be characterized using the word vocabulary as (s^z_1, s^z_2, ….,s^z_{Nz}); where $s^z_i \in W$ appears in the sequence (for i =1, 2, …, N_z). This representation of a word distinguishes between two or more clusters containing a common motif but differing in the exact number of occurrences in each of the clusters. Note that there is a word associated with a token not occurring within a gene sequence. Further, the model is invariant of the ordering of words within a sequence.

This model differs from the more commonly used Multivariate Bernoulli (MB) and Multinomial (MN) models frequently used in the context of document analysis [8], in the characterization of a sequence, and the way in which probabilities are assigned. In the MB model, a sequence is represented as a vector of binary attributes obtained as the presence or absence of tokens in the sequence from the token vocabulary. The number of times a word occurs within a sequence is not captured. Observe that the MB model is a special case of the motif model described here for R = 1. On the other hand, the MN model represents a sequence as a vector of token occurrences (frequency counts). Only the words that are present within a sequence are utilized for analyses, ignoring words that are absent. In comparison, we define here a motif model that imposes an upper bound on the frequency count, but takes non-occurring tokens into account. Both the MB and MN models assign a single probability value to each token irrespective of its frequency count, with the probability for non-occurrence calculated by subtracting from unity the probability of occurrence. This is in sharp contrast to our motif model, which allows assignment of any probability values to (token, frequency count) pairs, i.e., different frequency counts for the same token can have probabilities not restricted by any imposing model (the MN model imposes such restrictions). The design rationale behind our motif model is to capture meaningful frequency count information and also to treat tokens with different frequency counts independently to achieve high classification accuracy using a small number of words/tokens.

2.2 The Naïve Bayes Classifier

Let the K clusters be represented by the random variable C taking values $\{c_1, c_2,\ldots, c_K\}$, with prior probabilities denoted by $p(C_1)$, $p(C_2)$, …, $p(C_K)$. The Naïve Bayes model is *generative* in the sense that the word likelihood probabilities are first empirically estimated for each class using class prior probabilities and the dataset. It uses an easy to compute linear classifier to predict class membership for unknown sequences and has been shown to perform well in practice, especially suited to large datasets.

The likelihood probability of a word given a class, $p(w_i|C)$ based on the m-estimate [10] is estimated as $p(w_i \mid C_j) = (n_i^C + mp)/(|C| + m)$ where $n_i^{C_j}$ denotes the number of sequences in class C_j with word w_i occurring x_i times; $|C|$ is the total number of training examples in class C_j; p is the prior probability of words taken to be uniform = $1/|V|$;

and m is a constant called the 'equivalent sample size' which determines how heavily to weight p relative to the observed data. Imposing the Naïve Bayes conditional independence assumption, the conditional probability of $p(s^z|C_j)$ is a product of the individual probabilities $p(s^z \mid C_j) = p(s_1^z, s_2^z, ..., s_{N_z}^z \mid C_j) = \prod_i p(s_i^z \mid C_j)$. The posterior probabilities $p(C_j|s^z)$ can be calculated by a simple application of the Bayes theorem. Given a new sequence s^*, the classifier may be used in conjunction with the *maximum a posteriori* or *MAP* decision rule to assign it to a cluster.

$$classify(s^*) = \arg\max_j p(C_j) \prod_{i=1}^{|N_*|} p(s_i^* \mid C_j) \tag{2}$$

The independence assumption applied on the predictor variables, although not always accurate, simplifies the classification task dramatically by allowing the class conditional densities $p(x_k|C_j)$ to be calculated separately for each variable. In effect, Naive Bayes reduces a high-dimensional density estimation task to a one-dimensional kernel density estimation. The assumption does not seem to greatly affect the posterior probabilities, especially in regions near decision boundaries, thus leaving the classification task unaffected. If higher order interactions were to be examined, such as with Bayesian networks, the complete joint probability with size of vocabulary = $|V|*(R+1)$, $p(w_1, w_2, \ldots, w_n|Cj)$ has run-time exponential in the size of the vocabulary times the number of classes, $O(|S|^{|V|} * |C|)$. This is impractical for interactions of more than order 3, implying that it would be infeasible to identify CRMs with more than 3 combinatorial acting motifs. On the other hand, with the conditional independence assumption and assuming that the classification in independent of the positions of words (use same parameters for each position), the training time of the algorithm has order $O(|S|L_S + |C||V|)$ where L_S is the average length of each sequence. Even when the number of sequences examined is small, their lengths are usually at least 1kb or higher for eukaryotes and the number of unique words is quite high. Combining the two DNA strand orientations usually reduces the vocabulary size by $(1/3)^{rd}$, but is still higher than the number of sequences being examined. The test time for classification of new sequences takes $O(|C|L_t)$ time, where L_t is the average length of the test sequence.

2.3 Feature Selection to Identify Discriminative Cluster-Specific Motifs

The number of distinct words can be very large (in the order of $|V|*(R+l)$) depending on the number of genes and sequence lengths, the maximum stored token frequency and the motif length. Though good classification performance can often be achieved by considering all the words or even a large number (~1000) of words, it is not useful in understanding the regulatory mechanism underlying the differences in the gene clusters. Instead, we use a feature selection method to rank words by assigning scores to each word measuring its classification efficacy. The *Mutual Information MI* between two random variables measures the amount of information that the value of one variable gives about the value of the other [4] and has been shown to perform well in text

mining contexts [8]. $MI(X, Y) >= 0$ and equality holds if and only if the random variables X and Y are independent of each other. Discriminative motifs are selected from amongst the words that contribute significantly to the entire clustering. For each word $w \in W$, a new random variable $Y(w)=\{0,1\}$ is defined to represent the occurrence or non-occurrence of a word. The probabilities are estimated from the data as $p(y=1, C_j) = p(w, C_j)$ and $p(y=1) = p(w)$; while $p(y=0, C_j) = [1-p(w|C_j)]*p(C_j)$ and $p(y=0) = \sum_j p(y=0, C_j)$. The Mutual information between $Y(w)$ and C is then calculated as

$$MI(Y(w), C) = \sum_{j=1}^{K=|C|} \sum_{y=0,1} p(y, C_j) \log \frac{p(y, C_j)}{p(y) * p(C_j)} \tag{3}$$

2.4 Discriminative Algorithm

We now describe the discriminative algorithm that identifies cluster-specific motifs as the over-represented words within each cluster, that is words that are useful in discriminating a cluster from all the others. The mutual information feature selection criterion is used to score and sort all words in the vocabulary, the highest scores indicating the most informative words. The algorithm (described in Table 1) identifies discriminative motifs by iteratively dropping the least significant words (with zero score or below a specified threshold). Next, the relative score for each word, to the observed maximum score (maxScore), is calculated as [(maxScore – current word score) * 100/maxScore]%. Words with relative scores below a certain threshold of λ% (initialized to some λ_0) are removed from the complete word vocabulary W. If no word meets this condition indicating that the threshold is too stringent, the threshold is lowered by 5% and the removal step repeated. To avoid removing all the words in the vocabulary, user-defined value may be used to indicate the minimum number of words to be retained. The classifier is then re-trained using the new word vocabulary and the process repeated. The algorithm terminates when the desired number of discriminative motifs or a certain performance threshold has been reached.

Table 1. Algorithm to identify discriminative motifs

For each word w in the word vocabulary W,
(the occurrence of each word taken to be independent $\{0,1,\ldots,R\}$)
1. Get word counts from the data sequences.
2. Obtain estimates for likelihood and posteriors.
3. Calculate scores $MI(w_i, C_j)$.
4. Drop words with the lowest scores.
5. Use the remaining words to define the set of candidate motifs.
6. Classify using Naïve Bayes classifier.
7. Calculate classification accuracy (%) on a set of test sequences (with known classes).

Repeat steps 2 to 7 until error = (100-accuracy) < ε (small) or a small number of candidate features or motifs are obtained.

2.5 Constrained Enumeration Incorporating Sequence Conservation

The availability of diverse genomic databases allows the integration of these data into the motif finding process. Comparative genomics information has been used in some motif finding algorithms [1], [9], with the premise that functional regulatory elements are often conserved over evolution, thereby showing up in highly conserved regions of the alignments. These algorithms use as input all of the multiple sequence alignments to calculate some kind of scoring to measure the extent of conservation among the sequences. The calculations of these scores are data intensive and time consuming. DNA sequences do not change and can be considered a fixed data source, thereby eliminating the need to compute alignment scores dynamically within the motif finding algorithm. Instead, the same information can be accessed through pre-computed sequence conservation scores such as 'percent identity' or more comprehensive measures such as the PHYLOHMM scores [15]. To incorporate this quantitative measure into the algorithm, the following general procedure is followed.

Let Q denote any measure, such as conservation score for scoring the positions of a sequence, giving a higher score to regions of interest and a lower score to others. This information is incorporated into the motif finding process by considering sequence positions above a high preset threshold τ when scanning a sequence to extract tokens, $Q > \tau$. This constrains the enumeration of words, thereby influencing the likelihood probabilities based on the score information. The rest of the motif finding procedure remains the same.

2.6 Scalability and Implementation

Motif finding for higher eukaryotes like rodents and humans is a significantly greater challenge than for yeast or bacterial genomes in that transcription factor binding sites may be present at any distance from the transcription start site. Typically we will use a threshold of 3000bp to restrict the length of upstream and downstream sequences, but retain the complete intron sequences. The current implementation of the algorithm is designed to handle vast amounts of data - large number of genes and long sequences. The software is written in Java and connects to a MySQL database for input data and runtime storage. The PhyloHMM scores for the mouse genome require approximately 70GB of disk space. A typical run to enumerate about 360 genes with 4363 sequences (sequence lengths around 100kb) takes approximately 5 minutes without conservation and 15 minutes with conservation on an AMD Athlon XP 3000+ power with 1GB RAM.

3 Experimental Data and Results

3.1 Motif Assessment Data

The performance of our discriminative algorithm is assessed using benchmark data for mouse comparing the results with 13 other computational motif-finding tools [20]. The data consists of a total of 36 datasets with real binding sites from TRANSFAC planted within the sequences at their known positions and orientations, each dataset

consisting of one of three different types (12 datasets each) of background sequence namely (i)binding sites from real promoter sequences (called 'real'), (ii)randomly chosen promoter sequences from the same genome (called 'generic'), and (iii)sequences generated by a Markov chain of order 3 (called 'markov'). The prediction of only a single motif was allowed to be used in the comparisons for each dataset, using various statistical measures (*nSn, nPPV, nPC, nCC, sSn, sPPV,* and *sASP*) to assess the correctness of the predictions by comparing them with known binding sites (see [20] for more details). It was shown that the removal of the 'real' datasets resulted in an improvement in performance for nearly all the tools, with YMF being the most affected, while MotifSampler was the only tool to perform relatively better on the 'real' datasets than on the others, all species combined.

Discriminative motifs were predicted using the proposed algorithm using a set of background sequences generated from a third-order markov chain calculated with all the non-coding regions of the mouse genome as the second cluster. Three trials were performed for each dataset changing the background sequences. Fig. 2A summarizes the results obtained for all mouse data taken together and Fig. 2B compares the correlation measure *nCC* for each data type individually. Our method outperforms all the other tools on 'real' data with an *nCC* value of 0.114 (with nSN = 99) compared to the highest values of 0.1 and 0.08 (nSN = 50 and 34) achieved by MEME and MEME3 respectively and MotifSampler (nSN = 9) having a low negative value, indicating that it does not perform well on 'real' mouse data. Results on the 'generic' and 'markov' datasets could be poor due to the fact that the number of sequences within which to search for motifs is too few and the markov model probabilities were based on all noncoding sequences from the mouse genome rather than restricting it to use a maximum of 3000bp upstream of genes as was used to create the data. Hence, cluster discrimination is a better way to find motifs rather than relative over-representation over background, clearly sensitive to the model used to generate background sequences.

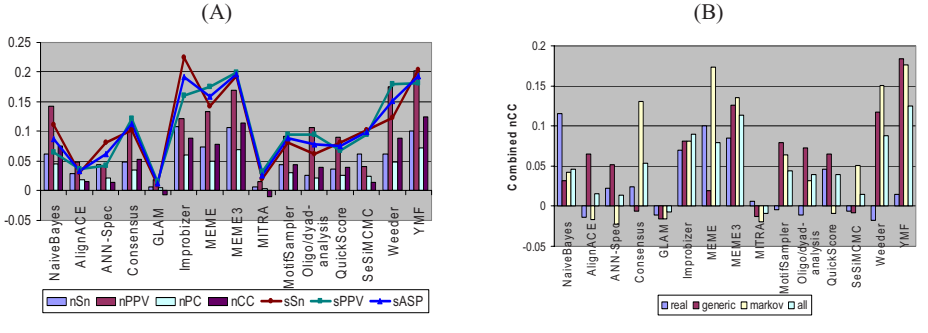

Fig. 2. Statistical measures of accuracy of the tools on the mouse benchmark data comparing 13 motif finding algorithms (MEME and MEME3 are considered one method with different parameters) (A) Combined measures of correctness over all 36 datasets for mouse. (B) Combined Correlation Coefficient (*nCC*) by data type.

3.2 Erythroid Differentiation in Mouse

The performance of the discriminative motif finding algorithm on real data is demonstrated using gene expression data of mouse genes studied with a late erythroid maturation model using the G1E line of Gata1-null cells, which are blocked at the proerythroblast stage because of the absence of the Gata1 transcription factor. The expression levels of over 9000 genes were measured at 6 time points after the restoration of Gata-1 [21], available via NCBI Gene Expression Omnibus database (accession GSE628). Genes were classified as being either up or down-regulated indicating that Gata-1 is important in gene repression as well as in the activation of a large number of genes. Two representative gene clusters (down and up-regulated) were picked after first clustering data using the kmeans algorithm with Pearson correlation distance to measure gene expression similarity. The chosen clusters are referred to as cluster1 and cluster2 respectively.

Many genes known to be induced by Gata-1 were present in the up-regulated group such as Alas2, Fog1, Vav2, Hbb-b1 and Hbb-b2. The Gata2 gene was down-regulated during maturation. The discriminative motif finding algorithm was used to identify transcription factors other than Gata-1, which may be responsible for the directional changes in expression levels. Repeat masked sequences from mm5 were used in the motif analysis comprised, taking as input all non-coding regions in and around genes (including UTRs and introns) and regions 3000bp flanking the gene upstream and downstream, gene positions as given by the KnownGenes table (UCSC table browser). Replicated genes (those for which the same Genbank Accession number was repeated in the KnownGenes table) were removed from the analysis and genes with overlapping positions were collapsed into a single gene complex. An examination of the distribution of sequence lengths showed a large amount of variability in the lengths of the introns, with the longest intron being approximately $12x10^4$ bp in length. It is important to note that existing motif finding tools cannot handle such large sequences.

The discriminative algorithm was used identify motifs specific to the up and down-regulated clusters, and classification performance compared for varying word lengths (l = 4, 6, and 8bp), and maximum word count, R from 1 to 10. Fig. 3 shows results obtained comparing classification accuracy with and without the use of conservation information using all words in W. For small values of R, shorter words (length l = 4) produce lowest classification accuracy ranging from 56% to 65% shown in Fig. 3A. Since the number of tokens occurring in the sequences (a sequence of length L has L-l+1 tokens) is large, there are many more common tokens between the sequences thereby producing less discrimination between the clusters. However as the word length increases (l=8) the number of tokens occurring in the sequences is fewer and classification accuracy is slightly better ranging from 60% to 65%. Although there are 4^8 possible distinct tokens that can be present we typically do not see so many in practice. The special case of our model when R=1 is the Multivariate Bernoulli model with lowest classification performance for all word lengths – 55% to 62%. Increasing the maximum word occurrence count, R from 10 to 50 resulted in a steady improvement in the classification performance of about 88% for any word length approximating the multinomial model, but resulting in an extremely large number of words $|V|*(l+1)$.

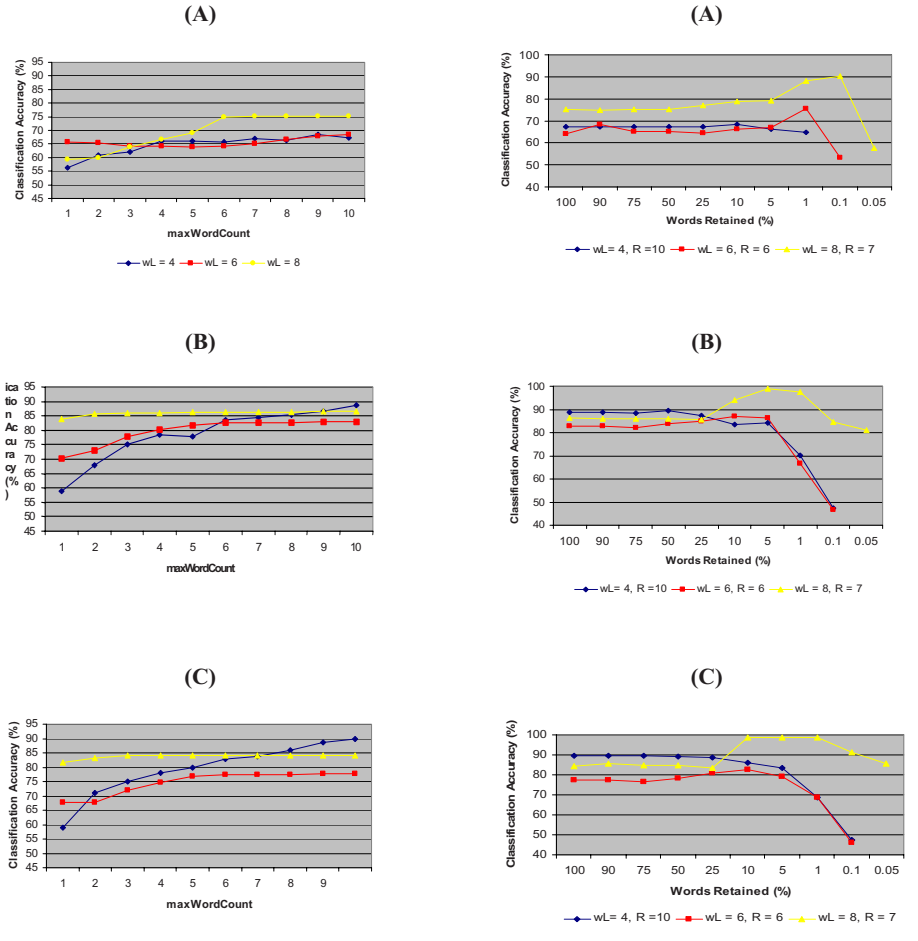

Fig. 3(left) and Fig. 4 (right). Classification accuracy of genes using the entire set of words (*W*) varying word length and maximum word count. (A)Unconstrained enumeration. (B) Constrained using conservation score threshold 0.7. (C) Constrained using conservation score threshold 0.8.

The use of multiple species sequence conservation in motif identification is evaluated next. We compare results with the constrained enumeration of words using thresholds 0.7 and 0.8 for PhyloHMM conservation scores for word lengths 4, 6 and 8 and varying *R* shown in Fig. 3B and C. A significant improvement by approximately 5% in classification accuracy is seen for word lengths 4 and 6, and nearly 20% for word length 8, for the worst case accuracy case (for *R*=1). As *R* increases, the accuracy is 90% for only *R*=10 when sequence conservation is used, while without conservation the maximum accuracy of 88% was achieved for *R*=50, clearly showing the sensitivity and performance quality of our word occurrence model. Even for a small word length of 4, the use of conservation scores achieves significant accuracy as the maximum word count increases. For word lengths 4 and 8 we observe that there is a

slight improvement in performance when a higher sequence conservation score threshold is used, while a word length of 6 has a reduction in accuracy from 84% to 78%. Using too few words for classification causes the reduction in accuracy. The same trend is observed when conservation information is used.

The effect of dropping words using the Mutual Information feature selection score is examined next. Fig. 4 shows that when words are dropped from 100% of the word vocabulary, W to 0.01% for a few word lengths and specific values of *R* the general trend is for the classification accuracy to increase significantly by steadily dropping words reaching a maximum value and then with a sharp drop when the number of words used for classification is further reduced. For example, for word length 8 and R=7, dropping 90% of the words (263160 distinct words) to 0.1% of words (263 distinct words) there is an improvement in accuracy of 15% from 75.28% to 90.28%.

Final motifs were returned for the top 50 scoring words. Typically, binding site sequences are variable, such that several words in the enumerative vocabulary may correspond to a single binding site. To consolidate the final list of motifs obtained and account for some of this variability, words differing in two or less positions were collapsed into a single unique motif (predicted binding site). For this, the similarity between every pair of words is required to be calculated, introducing the concept of a *distance* between words. The *Hamming distance* is one such measure of similarity between two strings, where it is calculated as "the number of positions in which the two strings differ, i.e., have different characters". This distance can be calculated only for strings of equal length. A more general and sophisticated measure of distance between strings is the *Levenshtein distance* and is defined for strings of arbitrary length. It counts the differences between two strings, where differences are counted not only when strings have different characters but also when one has a character whereas the other does not. The Levenshtein distance is what we use here to consolidate the final set of 50 words into a set of predicted motifs as listed here. Variable words are consolidated with other words only if they have equal word occurrence counts (described in the motif model). The consolidated motif is represented using the IUPAC nomenclature for nucleic acid representation.

The list of consolidated motifs is then matched against a list of known sites in human promoters [22] and shown in the following series of tables. Each consolidated motif is listed along with the motif for the matched known sites, the transcription factor to which it is known to bind, the word occurrence count, the corresponding mutual information score and the cluster for which the motif is discriminative. Table 2 lists the "most" discriminative motifs obtained using the discriminative algorithm with dropping words and without the use of conservation information. Tables 3 and 4 list the best set of motifs predicted when conservation information is incorporated into the feature selection model with conservation score (phylohmm) thresholds of 0.7 and 0.8 respectively. Conservation scores are based on mouse-centric multiple sequence alignments of human, mouse, rat and dog.

Several interesting aspects of the data emerge from the results of discriminative analysis. GATA1, which is a known transcription factor involved in erythroid differentiation and based on the experimental design, is most likely present in sequences of all genes irrespective of their cluster membership, is not identified. Clearly this binding site is not a candidate for discrimination between clusters and hence not detected by the algorithm. It can be observed that the integration of conservation information

Table 2. Discriminative Motifs without conservation

Factor	Known Site in Human	Predicted Site	WordCount	MI score	Discrimative for cluster
GABP	vCCGGAAGnGCR	CGGAAG	8	0.013196443	1
NCX	GTAAKTnG	GTAAAT	1	0.02081526	2
TBP	TATAAATW	GCTATAAA	5	0.0053224904	2
AREB6	WCAGGTGWnW, AbWCAGGTRnR	TCAGGTAA	5	0.0053224904	2
BACH2	SRTGAGTCAnC	TAGAGTCA	5	0.0053224904	2
NF-AT	WGGAAAnW	AGGAAA, CGGAAG	6, 8	0.010588359, 0.013196443	2, 1
CAC-BP	GRGGSTGGG	GGAGGTGG	6	0.008171844	2
LEF1	CTTTGA	CATTGA, CGTTGA	9, 6	0.0082450025, 0.007071697	1
MYC	SCACGTG	CATGTG	8	0.00951042	1
MYOD	RnCAGGTG	CATGTG	8	0.00951042	1
AP-4	GCAGCTGnY	CATGTG	8	0.00951042	1
SREBP-1	ATCACGTGAY	CATGTG	8	0.00951042	1
STAT5A,IY	AWTTCY, AWTTTCC	ATTTAC	8	0.02081526	2
AP-1, ATF-1	CTGASTCA, TGACGTCARRG	TAGAGTCA	5	0.0053224904	2
MAZ	GGGGAGGG	GGGAGGAT	4	0.0053224904	2
TAL-, ALPHA/E47	AACAGATGKT	CCAGATGT	5	0.0053224904	2

Table 3. Discriminative Motifs with conservation score threshold 0.7

Factor	Known Site in Human	Predicted Site	Word Count	MI score	Discrimative for cluster
NCX	GTAAKTnG	TGTAATTT	2	0.007647609	1
AP-4	GCAGCTGnY	CGAGCTGC	1	0.0117140515	1
		CAGCTG	2	0.0024077317	1
PU.1	WGAGGAAG	GAAGGAAG	2	0.019509021	2
FOXO1	RWAAACAA	CTAAACAG	1	0.0068486626	1
SF-1	TGRCCTTG	GACCTT	2	3.2390753E-6	2
MYOD	RnCAGGTG	GAGGTG	5	0.0024077317	1
AREB6	WCAGGTGWnW AbWCAGGTRnR	GAGGTG	5	0.0024077317	1

causes a smaller number of identified putative regulatory motifs to be matched with the list of “known binding sites” and with fewer word occurrence counts, when compared to the case where this data is not used in the analysis. It is generally believed that most “real” binding sites (TFBS) do not occur with high frequency within the gene’s non- coding sequence regions. It does appear to clearly indicate from our results that the addition of comparative genomics data causes a fewer number of false positives to be identified, keeping in mind the methods used to create the consensus sites and match with the “known sites” list.

Table 4. Discriminative Motifs with conservation score threshold 0.8

Factor	Known Site in Human	Predicted Site	Word Count	MI score	Discrimative for cluster
NCX	GTAAKTnG	GTAATTTT	1	0.010752671	1
		GTAATT	2	0.004678426	1
TBP	TATAAATW	GGTATAAA	1	0.009876572	1
AML1	ACCACA	ACCACA	3	0.0043065688	2
POU6F1	GCATAAWTTAT	ATTTAT	10	0.0049973438	1
DBP	GTdTGCT	TTTGCT	5	0.0047480455	2
CAC-BP	GRGGSTGGG	GGGTGG	4	0.0047480455	2
AP-4	GCAGCTGnY	CGAGCTGC	1	0.008132085	1
		CTAGCTGC	1	0.0072640753	1
NF-AT	WGGAAAnW	TGTAAA	10	0.0049973438	1
LEF1	CTTTGA	CTTTGT	4	0.0049973438	1
ER	RnnnTGACCT	GGACCT	2	0.004978211	1
STAT5A IY	AWTTCY AWTTTCC	GATTTC	1	0.00625898	2
TCF-4	WTCAAAGS	ACAAAG	4	0.0049973438	1
HNF-1	GGTTAATnWTTAMC	GTTA	6	0.003968242	1

There is only one predicted motif GTAAAT that matches the site for transcription factor NCX that is present in the list with a word occurrence count of 1 and is discriminative for the up-regulated cluster, cluster 2. Another NCX binding motif GTAATTTT is also identified when conservation information is used, but is however discriminative for the down-regulated cluster, cluster1. Literature indicates that NCX gene or *neural crest homeobox*, encodes a homeobox containing transcription factor that belongs to the Hox11 gene family, is involved in the activation of genes, and has been associated with diseases such as T-cell leukemia and Neuroblastoma. AP-4 or TFAP4 is another factor identified in all three results. This is a transcription factor of the basic helix-loop-helix-zipper (bHLH-ZIP) family contain a basic domain, which is used for DNA binding, and HLH and ZIP domains, which are used for oligomerization. Transcription factor AP4 activates both viral and cellular genes by binding to the symmetrical DNA sequence CAGCTG. Comparing the predicted motif within the three tables, it can be observed that the motifs CGAGCTGC and CTAGCTGC obtained when using conservation information are more similar to the known site in human with motif GCAGCTGnY when compared to that identified without the use of conservation scores – namely CATGTG. AP4 is also known to have functional interaction with AP-1 (a factor also identified by our algorithm).

Comparing results with and without sequence conservation, we see that factors AREB6 and MYOD are also identified at a conservation score threshold of 0.7, while TBP, NA-AT, and LEF1 are identified with a threshold of 0.8 implying higher degree of conservation over evolution. It is also interesting to see that some motifs that are discovered when conservation at a threshold value of 0.8 is used, such as AML1 and SF-1 do not appear in the motifs when conservation is not used. AML1 or runt-related transcription factor 1, RUNX1 is known to be associated with leukemia or as the name suggests, acute myeloid leukemia. The integration of conservation into the

analysis predicts many other interesting binding sites such as PU1, SF-1, and AML1 related to up-regulated genes while FOXO1, POU6F1, DBP, CAC-BP, ER, IY, TCF-4 and HNF-1 are associated with the down-regulated genes.

4 Conclusion

A new motif-finding algorithm is presented for multi-class discrimination. The motif model takes into account the frequency of token occurrence in individual sequences allowing for a very sensitive categorization of sequence clusters. Discriminative motifs are identified using an information-theoretic feature selection strategy and their prediction power examined with a supervised classifier. The algorithm is defined within a generic framework that allows the easy integration of additional genomic data, showing that comparative genomics information can be used to validate and evaluate the performance of the identified motifs. Results on benchmark and real data demonstrate the performance of our method in identifying true motifs. We have also provided strong empirical evidence to show that comparative genomics significantly improves the classification accuracy and achieves superior motif discovery.

References

1. Blanchette, M., Schwikowski, B., Tompa, M.: An exact algorithm to identify motifs in orthologous sequences from multiple species. In: Proc Eighth Intl Conf Intelligent Systems Mol Biol (ISMB), pp. 37–45. AAAI Press, Menlo Park (2000)
2. Bussemaker, H.J., Li, H., Siggia, E.D.: Regulatory element detection using correlation with expression. Nat. Gen. 27, 167–171 (2001)
3. Cardon, L., Stormo, G.: Expectation maximization for identifying protein-binding sites with variable lengths from unaligned DNA fragments. J. Mol. Biol. 223, 159–170 (1992)
4. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley, New York (1991)
5. Fickett, J.W., Wasserman, W.W.: Discovery and modeling of transcriptional regulatory regions. Curr. Opinion in Biotechnology 11, 19–24 (2000)
6. Holmes, I., Bruno, W.J.: Finding regulatory elements using joint likelihoods for sequence and expression profile data. Amer Assoc for Artificial Intelligence (2000)
7. Kasturi, J., Acharya, R.: Clustering of Diverse Genomic Data using Information Fusion. In: Proc ACM Sym Applied Computing (Bioinformatics Track) (2004)
8. McCallum, A., Nigam, K.: A comparison of event models for naïve bayes text classification. In: Proc. AAAI (1998)
9. McGuire, A.M., Church, G.M.: Predicting regulons and their cis-regulatory motifs by comparative genomics. Nucleic Acids Res. 28(22), 4523–4530 (2000)
10. Mitchell, T.: Machine Learning, ch. 10. McGraw Hill, New York (1997)
11. Lawrence, C., Reilly, A.: An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins 7, 41–51 (1990)
12. Liu, X., Brutlag, D.L., Liu, J.S.: Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. Pac. Sym. Biocomp., 27–38 (2001)
13. Roth, F.P., Hughes, J.D., Estep, P.W., Church, G.M.: Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nature Biotech. 16, 939–945 (1998)

14. Segal, E., Barash, Y., Simon, I., Friedman, N., Koller, D.: From promoter sequence to expression: a probabilistic framework. In: RECOMB (2001)
15. Siepel, A., Haussler, D.: Combining phylogenetic and hidden Markov models in biosequence analysis. In: Proc. Seventh Annual Intl. Conf. Comp. Mol. Biol. (RECOMB), pp. 277–286 (2003)
16. Sinha, S., Tompa, M.: A statistical method for finding transcription factor binding sites. Amer. Assoc. Artificial Intelligence (2000)
17. Sinha, S.: Discriminative motifs. J. Comput.Biol. 10(3-4), 599–615 (2003)
18. Smith, A.D., Sumazin, P., Zhang, M.Q.: Identifying tissue-selective transcription factor binding sites in vertebrate promoters. Proc. Natl. Acad. Sci. USA. 102(5), 1560–1565 (2005)
19. Thijs, G., Kathleen, M., Yves, M.: A Gibbs Sampling method to detect over-expressed motifs in the upstream regions of co-expressed genes. In: RECOMB (2001)
20. Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., Van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., Zhu, Z.: Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotech. 23(1), 137–144 (2005)
21. Welch, J.J., Watts, J.A., Vakoc, C.R., Yao, Y., Wang, H., Hardison, R.C., Blobel, G.A., Chodosh, L.A., Weiss, M.J.: Blood. 104(10), 3136–3147 (2004)
22. Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., Kellis, M.: Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434(7031), 338–345 (2005)