Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 13853, 13 pages
doi:10.1155/2007/13853

Research Article

Motif Discovery in Tissue-Specific Regulatory Sequences Using Directed Information

Arvind Rao,[1] Alfred O. Hero III,[1] David J. States,[2] and James Douglas Engel[3]

[1] *Departments of Electrical Engineering and Computer Science and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA*
[2] *Departments of Bioinformatics and Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA*
[3] *Department of Cell and Developmental Biology, University of Michigan, Ann Arbor, MI 48109, USA*

Received 1 March 2007; Revised 23 June 2007; Accepted 17 September 2007

Recommended by Teemu Roos

Motif discovery for the identification of functional regulatory elements underlying gene expression is a challenging problem. Sequence inspection often leads to discovery of novel motifs (including transcription factor sites) with previously uncharacterized function in gene expression. Coupled with the complexity underlying tissue-specific gene expression, there are several motifs that are putatively responsible for expression in a certain cell type. This has important implications in understanding fundamental biological processes such as development and disease progression. In this work, we present an approach to the identification of motifs (not necessarily transcription factor sites) and examine its application to some questions in current bioinformatics research. These motifs are seen to discriminate tissue-specific gene promoter or regulatory regions from those that are not tissue-specific. There are two main contributions of this work. Firstly, we propose the use of directed information for such classification constrained motif discovery, and then use the selected features with a support vector machine (SVM) classifier to find the tissue specificity of any sequence of interest. Such analysis yields several novel interesting motifs that merit further experimental characterization. Furthermore, this approach leads to a principled framework for the prospective examination of any chosen motif to be discriminatory motif for a group of coexpressed/coregulated genes, thereby integrating sequence and expression perspectives. We hypothesize that the discovery of these motifs would enable the large-scale investigation for the tissue-specific regulatory role of any conserved sequence element identified from genome-wide studies.

1. INTRODUCTION

Understanding the mechanisms underlying regulation of tissue-specific gene expression remains a challenging question. While all mature cells in the body have a complete copy of the human genome, each cell type only expresses those genes it needs to carry out its assigned task. This includes genes required for basic cellular maintenance (often called "housekeeping genes") and those genes whose function is specific to the particular tissue type that the cell belongs to. Gene expression by a way of transcription is the process of generation of messenger RNA (mRNA) from the DNA template representing the gene. It is the intermediate step before the generation of functional protein from messenger RNA. During gene expression (see Figure 1), transcription factor (TF) proteins are recruited at the proximal promoter of the gene as well as at sequence elements (enhancers/silencers) which can lie several hundreds of kilobases from the gene's transcriptional start site (TSS). The basal transcriptional machinery at the promoter coupled with the transcription factor complexes at these distal, long-range regulatory elements (LREs) are collectively involved in directing tissue-specific expression of genes.

One of the current challenges in the post-genomic era is the principled discovery of such LREs genome-wide. Recently, there has been a community-wide effort (http://www.genome.gov/ENCODE) to find all regulatory elements in 1% of the human genome. The examination of the discovered elements would reveal characteristics typical of most enhancers which would aid their principled discovery and examination on a genome-wide scale. Some characteristics of experimentally identified distal regulatory elements [1, 2] are as follows.

(i) Noncoding elements: distal regulatory elements are noncoding and can either be intronic or intergenic regions on the genome. Hence, previous models for gene

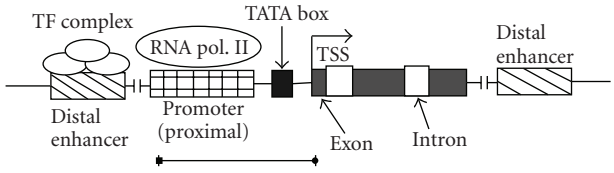

Figure 1: Schematic of transcriptional regulation. Sequence motifs at the promoter and the distal regulatory elements together confer specificity of gene expression via TF binding.

finding [3] are not directly applicable. With over 98% of the annotated genome being noncoding, the precise localization of regulatory elements that underlie tissue-specific gene expression is a challenging problem.

(ii) Distance/orientation independent: an enhancer can act from variable genomic distances (hundreds of kilobases) to regulate gene expression in conjunction with the proximal promoter, possibly via a looping mechanism [4]. These enhancers can lie upstream or downstream of the actual gene along the genomic locus.

(iii) Promoter dependent: since the action at a distance of these elements involves the recruitment of TFs that direct tissue-specific gene expression, the promoter that they interact with is critical.

Although there are instances where a gene harbors tissue-specific activity at the promoter itself, the role of long-range elements (LREs) remains of interest, for example, for a detailed understanding of their regulatory role in gene expression during biological processes like organ development and disease progression [5]. We seek to develop computational strategies to find novel LREs genome-wide that govern tissue specific expression for any gene of interest. A common approach for their discovery is the use of motif-based sequence signatures. Any sequence element can then be scanned for such a signature and its tissue specificity can be ascertained [6].

Thus, our primary question in this regard is that is there a discriminating sequence property of LRE elements that determines tissue-specific gene expression—more particularly, are there any sequence motifs in known regulatory elements that can aid discovery of new elements [7]. To answer this, we examine known tissue-specific regulatory elements (promoters and enhancers) for motifs that discriminate them from a background set of neutral elements (such as housekeeping gene promoters). For this study, the datasets are derived from the following sources.

(i) *Promoters of tissue-specific genes*: before the widespread discovery of long-range regulatory elements (LREs), it was hypothesized that promoters governed gene expression alone. There is substantial evidence for the binding of tissue-specific transcription factors at the promoters of expressed genes. This suggests that in spite of newer information implicating the role of LREs, promoters also have interesting motifs that govern tissue-specific expression.

Another practical reason for the examination of promoters is that their locations (and genomic sequences) are more clearly delineated on genome databases (like UCSC or Ensembl). Sufficient data (http://symatlas.gnf.org) on the expression of genes is also publicly available for analysis. Sequence motif discovery is set up as a feature extraction problem from these tissue-specific promoter sequences. Subsequently, a support vector machine (SVM) classifier is used to classify new promoters into specific and nonspecific categories based on the identified sequence features (motifs). Using the SVM classifier algorithm, 90% of tissue-specific genes are correctly classified based upon their upstream promoter region sequences alone.

(ii) *Known long range regulatory elements (LRE) motifs*: to analyze the motifs in LRE elements, we examine the results of the above approach on the Enhancer Browser dataset (http://enhancer.lbl.gov) which has results of expression of ultraconserved genomic elements in transgenic mice [8]. An examination of these ultraconserved enhancers is useful for the extraction of discriminatory motifs to distinguish the regulatory elements from the nonregulatory (neutral) ones. Here the results indicate that up to 95% of the sequences can be correctly classified using these identified motifs.

We note that some of the identified motifs might not be transcription factor binding motifs, and would need to be functionally characterized. This is an advantage of our method-instead of constraining ourselves to the degeneracy present in TF databases (like TRANSFAC/JASPAR), we look for all sequences of a fixed length.

2. CONTRIBUTIONS

Using microarray gene expression data, [9, 10] proposes an approach to assign genes into tissue-specific and nonspecific categories using an entropy criterion. Variation in expression and its divergence from ubiquitous expression (uniform distribution across all tissue types) is used to make this assignment. Based on such assignment, several features like CpG island density, frequency of transcription factor motif occurrence, can be examined to potentially discriminate these two groups. Other work has explored the existence of key motifs (transcription factor binding sites) in the promoters of tissue-specific genes (see [11, 12]). Based on the successes reported in these methods, it is expected that a principled examination and characterization of every sequence motif identified to be discriminatory might lead to improved insight into the biology of gene regulation. For example, such a strategy might lead to the discovery of newer TFBS motifs, as well as those underlying epigenetic phenomena.

For the purpose of identifying discriminative motifs from the training data (tissue-specific promoters or LREs), our approach is as follows.

(i) *Variable selection*: firstly, sequence motifs that discriminate between tissue-specific and non-specific elements are discovered. In machine learning, this is a feature selection problem with features being the

counts of sequence motifs in the training sequences. Without loss of generality, six-nucleotide motifs (hexamers) are used as motif features. This is based on the observation that most transcription factor binding motifs have a 5-6 nucleotide core sequence with degeneracy at the ends of the motif. A similar setup has been introduced in [13–15]. The motif search space is, therefore, a $4^6 = 4096$-dimensional one. The presented approach, however, does not depend on motif length and can be scaled according to biological knowledge. For variable (motif) selection, a novel feature selection approach (based on an information theoretic quantity called *directed information* (DI)) is proposed. The improved performance of this criterion over using mutual information for motif selection is also demonstrated.

(ii) *Classifier design*: after discovering discriminating motifs using the above DI step, an SVM classifier that separates the samples between the two classes (specific and nonspecific) from this motif space is constructed.

Apart from this novel feature selection approach, several questions pertaining to bioinformatics methodology can be potentially answered using this framework—some of these are as follows.

(i) Are there common motifs underlying tissue-specific expression that are identified from tissue-specific promoters and enhancers? In this paper, an examination of motifs (from promoters and enhancers) corresponding to brain-specific expression is done to address this question.

(ii) Do these motifs correspond to known motifs (transcription factor binding sites)? We show that several motifs are indeed consensus sites for transcription factor binding, although their real role can only be identified in conjunction with experimental evidence.

(iii) Is it possible to relate the motif information from the sequence and expression perspectives to understand regulatory mechanisms? This question is addressed in Section 11.3.

(iv) How useful are these motifs in predicting new tissue-specific regulatory elements? This is partly explained from the results of SVM classification.

This work differs from that in [13, 14], in several aspects. We present the DI-based feature selection procedure as part of an overall unified framework to answer several questions in bioinformatics, not limited to finding discriminating motifs between two classes of sequences. Particularly, one of the advantages is the ability to examine any particular motif as a potential discriminator between two classes. Also, this work accounts for the notion of tissue-specificity of promoters/enhancers (in line with more recent work in [8–10, 16, 17]). Also, this framework enables the principled integration of various data sources to address the above questions. These are clarified in Section 11.

3. RATIONALE

The main approaches to finding common motifs driving tissue-specific gene regulation are summarized in [1, 2]. The most common approach is to look for TFBS motifs that are statistically over-represented in the promoters of the coexpressed genes based on a background (binomial or Poisson) distribution of motif occurrence genomewide.

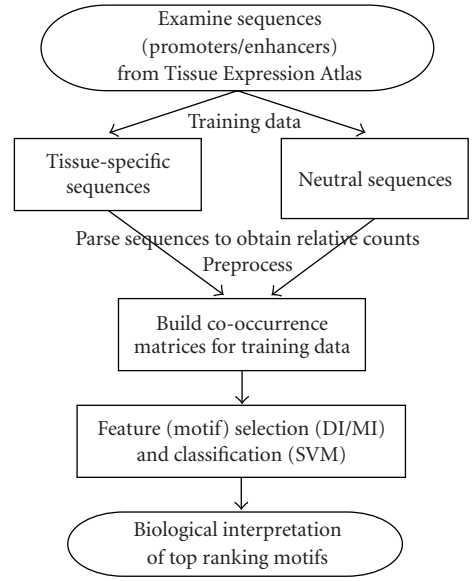

Figure 2: An overview of the proposed approach. Each of the steps are outlined in the following sections.

In this work, the problem of motif discovery is set up as follows. Using two annotated groups of genes, tissue-specific ("*ts*") and nontissue-specific ("*nts*"), hexamer motifs that best discriminate these two classes are found. The goal would be to make this set of motifs as small as possible, that is, to achieve maximal class partitioning with the smallest feature subset.

Several metrics have been proposed to find features with maximal class label association. From information theory, mutual information is a popular choice [18]. This is a symmetric association metric and does not resolve the direction of dependency (i.e., if features depend on the class label or vice versa). It is important to find features that induce the class label. Feature selection from data implies selection (control) of a feature subset that maximally captures the underlying character (class label) of the data. There is no control over the label (a purely observational characterization).

With this motivation, a new metric for discriminative hexamer subset selection, termed "directed information" (DI), is proposed. Based on the selected features, a classifier is used to classify sequences to tissue-specific or nontissue-specific categories. The performance of this DI-based feature selection metric is subsequently evaluated in the context of the SVM classifier.

4. OVERALL METHODOLOGY

The overall schematic of the proposed procedure is outlined in Figure 2.

Below we present our approach to find promoter-specific or enhancer-specific motifs.

5. MOTIF ACQUISITION

5.1. Promoter motifs

5.1.1. Microarray analysis

Raw microarray data is available from the Novartis Foundation (GNF) [http://symatlas.gnf.org]. Data is normalized using RMA from the bioconductor packages for R [http://cran.r-project.org]. Following normalization, replicate samples are averaged together. Only 25 tissue types are used in our analysis including: adrenal gland, amygdala, brain, caudate nucleus, cerebellum, corpus callosum, cortex, dorsal root ganglion, heart, HUVEC, kidney, liver, lung, pancreas, pituitary, placenta, salivary, spinal cord, spleen, testis, thalamus, thymus, thyroid, trachea, and uterus.

In this context, the notion of tissue specificity of a gene needs clarification. Suppose there are N genes, $g_1, g_2, \ldots, g_N$, and T tissue types (in GNF: $T = 25$), we construct an $N \times T$ tissue specificity matrix: $M = [0]_{N\times T}$. For each gene $g_i, 1 \le i \le N$, let $g_{i,[0.5T]} = \text{median}(g_{i,k})$, for all $k \in 1, 2, \ldots, T$; $g_{i,k}$ being the expression level of gene i in tissue k. Define each entry $M_{i,k}$ as

$$M_{i,k} = \begin{cases} 1 & \text{if } g_{i,k} \ge 2g_{i,[0.5T]}, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Now consider the N-dimensional vector $m_i = \sum_{k=1}^{T} M_{i,k}, 1 \le i \le N$, that is, summing all the columns of each row. The interquartile range of $'m'$ can be used for "*ts*"/"*nts*" assignment. Gene indices $'i'$ that are in quartile 1 (= 3) are labeled as "*ts*," and those in quartile 4 (= 22) are labeled as "*nts*."

With this approach, a total of 1924 probes representing 1817 genes were classified as tissue-specific, while 2006 probes representing 2273 genes were classified as nontissue-specific. In this work, genes which are either heart-specific or brain-specific are considered. From the tissue-specific genes obtained from the above approach, 45 brain-specific gene promoters and 118 heart-specific gene promoters are obtained. As mentioned in Section 2, one of the objectives is to find motifs that are responsible for brain/heart specific expression and also correlate them with binding profiles of known transcription factor binding motifs.

5.1.2. Sequence analysis

Genes ("*ts*" or "*nts*") associated with candidate probes are identified using the Ensembl Ensmart [http://www.ensembl.org] tool. For each gene, sequence from 2000 bp upstream and 1000 bp down-stream upto the start of the first exon relative to their reported TSS is extracted from the Ensembl Genome Database (Release 37). The relative counts of each of the 4^6 hexamers are computed within each gene promoter sequence of the two categories ("ts" and "nts")—using the "*seqinr*" library in the R environment. A t-test is performed between the relative counts of each hexamer between the two expression categories ("*ts*" and "*nts*") and the top 1000 significant hexamers ($\vec{\mathbf{H}} = H_1, H_2, \ldots, H_{1000}$) are obtained. The relative counts of these hexamers is recomputed for each gene individually. This results in two hexamer-gene cooccurrence matrices—one for the "*ts*" class (dimension $N_{\text{train},+1} \times 1000$) and the other for the "*nts*" class (dimension $N_{\text{train},-1} \times 1000$). Here $N_{\text{train},+1}$ and $N_{\text{train},-1}$ are the number of positive training and negative training samples, respectively.

Table 1: The "motif frequency matrix" for a set of gene promoters. The first column is their ENSEMBL gene identifiers and the other 4 columns are the motifs. A cell entry denotes the number of times a given motif occurs in the upstream (−2000 to +1000 bp from TSS) region of each corresponding gene.

Ensembl Gene ID	AAAAAA	AAAAAG	AAAAAT	AAAACA
ENSG00000155366	0	0	1	4
ENSG00001780892	6	5	5	6
ENSG00000189171	1	2	1	0
ENSG00000168664	6	3	8	0
ENSG00000160917	4	1	4	2
ENSG00000163655	2	4	0	1
ENSG00001228844	8	6	10	7
ENSG00000176749	0	0	0	0
ENSG00000006451	5	2	2	1

The input to the feature selection procedure is a gene promoter-motif frequency table (Table 1). The genes relevant to each class are identified from tissue microarray analysis, following steps in Section 5.1.1 and the frequency table is built by parsing the gene promoters for the presence of each of the $4^6 = 4096$ possible hexamers.

5.2. LRE motifs

To analyze long range elements which confer tissue-specific expression, the Mouse Enhancer database (http://enhancer.lbl.gov) is examined. This database has a list of experimentally validated ultraconserved elements which have been tested for tissue specific expression in transgenic mice [8], and can be searched for a list of all elements which have expression in a tissue of interest. In this work, we consider expression in tissues relating to the developing brain. According to the experimental protocol, the various regions are cloned upstream of a heat shock protein promoter (*hsp68-lacz*), thereby not adhering to the idea of promoter specificity in tissue-specific expression. Though this is of concern in that there is loss of some gene-specific information, we work with this data since we are more interested in tissue expression and also due to a paucity of public promoter-dependent enhancer data.

This database also has a collection of ultraconserved elements that do not have any transgenic expression in vivo. This is used as the neutral/background set of data which corresponds to the "*nts*" (nontissue-specific class) for feature selection and classifier design.

As in the above (promoter) case, these sequences (seventy four enhancers for brain-specific expression) are parsed for the absolute counts of the 4096 hexamers, a cooccurrence matrix ($N_{\text{train},+1} = 74$) is built and then t-test P-values are used to find the top 1000 hexamers ($\vec{\mathbf{H}}' = H'_1, H'_2, \ldots, H'_{1000}$)

that are maximally different between the two classes (brain-specific and brain-nonspecific).

The next three sections clarify the preprocessing, feature selection, and classifier design steps to mine these cooccurrence matrices for hexamer motifs that are strongly associated with the class label. We note that though this work is illustrated using two class labels, the approach can be extended in a straightforward way to the multiclass problem.

6. PREPROCESSING

From the above, $N_{\text{train},+1} \times 1000$ and $N_{\text{train},-1} \times 1000$ dimensional cooccurrence matrices are available for the tissue-specific and nonspecific data, both for the promoter and enhancer sequences. Before proceeding to the feature (hexamer motif) selection step, the counts of the $M = 1000$ hexamers in each training sample need to be normalized to account for variable sequence lengths. In the cooccurrence matrix, let $gc_{i,k}$ represent the absolute count of the kth hexamer, $k \in 1, 2, \ldots, M$, in the ith gene. Then, for each gene g_i, the quantile labeled matrix has $X_{i,k} = l$ if $gc_{i,[((l-1)/K)M]} \leq gc_{i,k} < gc_{i,[(l/K)M]}, K = 4$. Matrices of dimension $N_{\text{train},+1} \times 1001$, $N_{\text{train},-1} \times 1001$ for the specific and nonspecific training samples are now obtained. Each matrix contains the quantile label assignments for the 1000 hexamers ($X_i, i \in (1, 2, \ldots, 1000)$), as stated above, and the last column has the corresponding class label ($Y = -1/+1$).

7. DIRECTED INFORMATION AND FEATURE SELECTION

The primary goal in feature selection is to find the minimal subset of features (from hexamers: $\vec{\mathbf{H}}/\vec{\mathbf{H}}'$) that lead to maximal discrimination of the class label ($Y_i \in (-1/+1)$), using each of the $i \in (1, 2, \ldots, (N_{\text{train},+1} + N_{\text{train},-1}))$ genes during training. We are looking for a subset of the variables $(H_{i,1}, \ldots, H_{i,1000})$ which are directionally associated with the class label (Y_i). These hexamers putatively influence/induce the class label (see Figure 3). As can be seen from [19], there is considerable interest in discovering such dependencies from expression and sequence data. Following [20], we search for features (in *measurement* space) that induce the class label (in *observation* space).

One way to interpret the feature selection problem is the following: nature is trying to communicate a source symbol ($Y \in \{-1/+1\}$), corresponding to the gene class label ("*nts/ts*"), to us. In this setup, an encoder that extracts frequencies of a particular hexamer (H_i) maps the source symbol (Y) to $H_i(Y)$. The decoder outputs the source reconstruction $\hat{Y}$ based on the received codeword $c_i(Y) = H_i(Y)$.

We observe that there are several possible encoding schemes $c_i(Y)$ that the encoder could potentially use ($i = 1, 2, \ldots, 1000$), each corresponding to feature extraction via a different hexamer H_i. An encoder is the mapping rule $c_i : Y \rightarrow H_i$. The ideal encoding scheme is one which induces the most discriminative partitioning of the code (feature) space, for successful reconstruction of Y by the decoder. The ranking of each encoder's performance over all possible mappings yields the most discriminative mapping. This measure

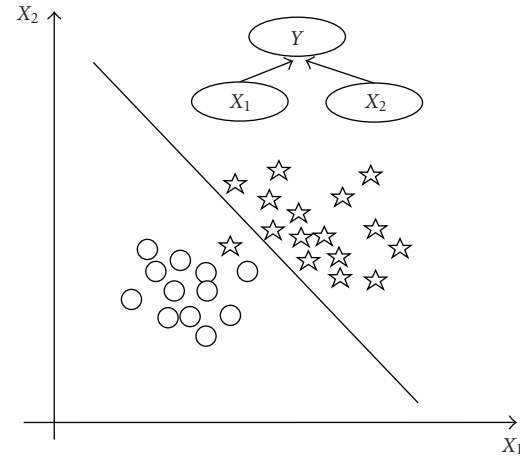

Figure 3: Causal feature discovery for two class discrimination, adapted from [20]. Here the variables X_1 and X_2 discriminate Y, the class label.

of performance is the amount of information flow from the mapping (hexamer) to the class label. Using mutual information as one such measure indeed identifies the best features [18], but fails to resolve the direction of dependence due to its symmetric nature $I(H_i; Y) = I(Y; H_i)$. The direction of dependence is important since it pinpoints those features that induce the class label (not vice versa). This is necessary since these class labels are predetermined (given to us by biology) and the only control we have is the feature space onto which we project the data points, for the purpose of classification. This loosely parallels the use the directed edges in Bayesian networks for inference of feature-class label associations [20].

Unlike mutual information (MI), directed information (DI) is a metric to quantify the directed flow of information. It was originally introduced in [21, 22] to examine the transfer of information from encoder to decoder under feedback/feedforward scenarios and to resolve directivity during bidirectional information transfer. Given its utility in the encoding of sources with memory (correlated sources), this work demonstrates it to be a competitive metric to MI for feature selection in learning problems. DI answers which of the encoding schemes (corresponding to each hexamer H_i) leads to maximal information transfer from the hexamer labels to the class labels (i.e., directed dependency).

The DI is a measure of the directed dependence between two vectors $X_i = [X_{1,i}, X_{2,i}, \ldots, X_{n,i}]$ and $Y = [Y_1, Y_2, \ldots, Y_n]$. In this case, $X_{j,i}$ = quantile label for the frequency of hexamer $i \in (1, 2, \ldots, 1000)$ in the jth training sequence. $Y = [Y_1, Y_2, \ldots, Y_n]$ are the corresponding class labels $(-1, +1)$. For a block length N, the DI is given by [22]

$$I(X_i^N \longrightarrow Y^N) = \sum_{n=1}^{N} I(X_i^n; Y_n \mid Y^{n-1}). \tag{2}$$

Using a stationarity assumption over a finite-length memory of the training samples, a correspondence with the setup in [22, 23] can be seen. As already known [24], the mutual information is $I(X^N; Y^N) = H(X^N) - H(X^N \mid Y^N)$, where $H(X^N)$ and $H(X^N \mid Y^N)$ are the Shannon entropy of X^N and

the conditional entropy of X^N given Y^N, respectively. With this definition of mutual information, the directed information simplifies to

$$\begin{aligned} I(X^N \longrightarrow Y^N) &= \sum_{n=1}^{N} \left[H(X^n \mid Y^{n-1}) - H(X^n \mid Y^n)\right] \\ &= \sum_{n=1}^{N} \Big\{\left[H(X^n, Y^{n-1}) - H(Y^{n-1})\right] \\ &\qquad - \left[H(X^n, Y^n) - H(Y^n)\right]\Big\}. \end{aligned} \tag{3}$$

Using (3), the directed information is expressed in terms of individual and joint entropies of X^n and Y^n. This expression implies the need for higher-order entropy estimation from a moderate sample size. A Voronoi-tessellation-based [25] adaptive partitioning of the observation space can handle $N = 5/6$ without much complexity.

The relationship between MI and DI is given by [22] DI: $I(X^N \to Y^N) = \sum_{i=1}^{N} I(X^i; Y_i \mid Y^{i-1})$,

MI: $I(X^N; Y^N) = \sum_{i=1}^{N} I(X^N; Y_i \mid Y^{i-1}) = I(X^N \to Y^N) + I(0Y^{N-1} \to X^N)$.

To clarify, $I(X^N \to Y^N)$ is the directed information from X to Y, whereas $I(0Y^{N-1} \to X^N)$ is the directed information from a (one-sample) delayed version of Y^N to X^N. From [23], it is clear that DI resolves the direction of information transfer (feedback or feedforward). If there is no feedback/feedforward, $I(X^N \to Y^N) = I(X^N; Y^N)$.

From the above chain-rule formulations for DI and MI, it is clear that the expression for DI is permutation-variant (i.e., the value of the DI is different for a different ordering of random variables). Thus, we instead find the $I_p(X^N \to Y^N)$, a DI measure for a particular ordering of the N random variables (r.v.'s). The DI value for our purpose, $I(X^N \to Y^N)$ is an average over all possible sample permutations given by $I(X^N \to Y^N) = (1/N!) \sum_{p=1}^{N!} I_p(X^N \to Y^N)$. For MI, however, $I_p(X^N; Y^N) = I(X^N; Y^N)$, because MI is permutation-invariant (i.e., independent of r.v.'s ordering). As can be readily observed, this problem is combinatorially complex, and hence, a Monte Carlo sampling strategy (1000 trials) is used for computing $I(X^N \to Y^N)$. This is because we find that about 1000 trials yields a DI confidence interval (CI) that is only 20% more than the corresponding CI obtained from 10000 trials of the data, a far more exhaustive number.

To select features, we maximize $I(X^N \to Y^N)$ over the possible pairs $(\vec{\mathbf{X}}, Y)$. This feature selection problem for the ith training instance reduces to identifying which hexamer ($k \in (1, 2, \ldots, 4096)$) has the highest $I(X_k \to Y)$.

The higher-dimensional entropy can be estimated using order statistics of the observed samples [25] by iterative partitioning of the observation space until nearly uniform partitions are obtained. This method lends itself to a partitioning scheme that can be used for entropy estimation even for a moderate number of samples in the observation space of the underlying probability distribution. Several such algorithms for adaptive density estimation have been proposed (see [26–28]) and can find potential application in this procedure. In this methodology, a Voronoi tessellation approach for entropy estimation because of the higher performance guarantees as well as the relative ease of implementation of such a procedure.

The above method is used to estimate the true DI between a given hexamer and the class label for the entire training set. Feature selection comprises of finding all those hexamers (X_i) for which $I(X_i^N \to Y^N)$ is the highest. From the definition of DI, we know that $0 \le I(X_i^N \to Y^N) \le I(X_i^N; Y^N) < \infty$. To make a meaningful comparison of the strengths of association between different hexamers and the class label, we use a normalized score to rank the DI values. This normalized measure ρ_{DI} should be able to map this large range ($[0, \infty]$) to $[0, 1]$. Following [29], an expression for the normalized DI is given by

$$\begin{aligned} \rho_{\mathrm{DI}} &= \sqrt{1 - e^{-2I(X^N \to Y^N)}} \\ &= \sqrt{1 - e^{-2\sum_{i=1}^{N} I(X^i; Y_i \mid Y^{i-1})}}. \end{aligned} \tag{4}$$

Another point of consideration is to estimate the significance of the DI value compared to a null distribution on the DI value (i.e., what is the chance of finding the DI value by chance from the N-length series X_i and Y). This is done using confidence intervals after permutation testing (Section 8).

8. BOOTSTRAPPED CONFIDENCE INTERVALS

In the absence of knowledge of the true distribution of the DI estimate, an approximate confidence interval for the DI estimate ($\hat{I}(X^N \to Y^N)$) is found using bootstrapping [30]. Density estimation is based on kernel smoothing over the bootstrapped samples [31].

The kernel density estimate for the bootstrapped DI (with $n = 1000$ samples), $Z \triangleq \hat{I}_B(X^N \to Y^N)$ becomes $\hat{f}_h(Z) = (1/nh) \sum_{i=1}^{n} (3/4)[1 - ((z_i - z)/h)^2] I(|(z_i - z)/h| \le 1)$ with $h \approx 2.67\hat{\sigma}_z$ and $n = 1000$. $\hat{I}_B(X^N \to Y^N)$ is obtained by finding the DI for each random permutation of the X, Y series, and performing this permutation B times. As it is clear from the above expression, the Epanechnikov kernel is used for density estimation from the bootstrapped samples. The choice of the kernel is based on its excellent characteristics—a compact region of support, the lowest asymptotic mean squared error (AMISE) and favorable bias-variance tradeoff [31].

We denote the cumulative distribution function (over the bootstrap samples) of $\hat{I}(X^N \to Y^N)$ by $F_{\hat{I}_B(X^N \to Y^N)}(\hat{I}_B(X^N \to Y^N))$. Let the mean of the bootstrapped null distribution be $I_B^*(X^N \to Y^N)$. We denote by $t_{1-\alpha}$, the $(1 - \alpha)$th quantile of this distribution, that is, $\{t_{1-\alpha} : P([(\hat{I}_B(X^N \to Y^N) - I_B^*(X^N \to Y^N))/\hat{\sigma}] \le t_{1-\alpha}) = 1 - \alpha\}$. Since we need the true $\hat{I}(X^N \to Y^N)$ to be significant and close to 1, we need $\hat{I}(X^N \to Y^N) \ge [I_B^*(X^N \to Y^N) + t_{1-\alpha} \times \hat{\sigma}]$, with $\hat{\sigma}$ being the standard error of the bootstrapped distribution, $\hat{\sigma} = \sqrt{([\Sigma_{b=1}^{B} \hat{I}_b(X^N \to Y^N) - I_B^*(X^N \to Y^N)]^2)/(B-1)}$; B is the number of bootstrap samples.

This hypothesis test is done for each of the 1000 motifs, in order to select the top $'d'$ motifs based on DI value, which is then used for classifier training subsequently. This leads to a need for multiple-testing correction. Because the Bonferroni correction is extremely stringent in such settings, the Benjamini-Hochberg procedure [32], which has a higher false positive rate but a lower false negative rate, is used in this work.

9. SUPPORT VECTOR MACHINES

From the top d features identified from the ranked list of features having high DI with the class label, a support vector machine classifier in these d dimensions is designed. An SVM is a hyperplane classifier which operates by finding a maximum margin linear hyperplane to separate two different classes of data in high-dimensional ($D > d$) space. The training data has $N(= N_{\text{train},+1} + N_{\text{train},-1})$ pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, with $x_i \in \mathcal{R}^d$ and $y_i \in \{-1, +1\}$.

An SVM is a maximum margin hyperplane classifier in a nonlinearly extended high-dimensional space. For extending the dimensions from d to $D > d$, a radial basis kernel is used.

The objective is to minimize $\|\beta\|$ in the hyperplane $\{x : f(x) = x^T\beta + \beta_0\}$, subject to $y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \, \forall i, \xi_i \geq 0, \sum \xi_i \leq$ constant [33].

10. SUMMARY OF OVERALL APPROACH

Our proposed approach is as follows. Here, the term "sequence" can pertain to either tissue-specific promoters or LRE sequences, obtained from the GNF SymAtlas and Ensembl databases or the Enhancer Browser.

(1) The sequence is parsed to obtain the relative counts/frequencies of occurrence of the hexamer in that sequence and to build the hexamer-sequence frequency matrix. The "*seqinr*" package in R is used for this purpose. This is done for all the sequences in the specific (class "+1") and nonspecific (class "−1") categories. The matrix thus has $N = N_{\text{train},+1} + N_{\text{train},-1}$ rows and $4^6 = 4096$ columns.

(2) The obtained hexamer-sequence frequency matrix is preprocessed by assigning quantile labels for each hexamer within the ith sequence. A hexamer-sequence matrix is thus obtained where the (i, j)th entry has the quantile label of the jth hexamer in the ith sequence. This is done for all the N training sequences consisting of examples from the −1 and +1 class labels.

(3) Thus, two submatrices corresponding to the two class labels are built. One matrix contains the hexamer-sequence quantile labels for the positive training examples and the other matrix is for the negative training examples.

(4) To select hexamers that are most different between the positive and negative training examples, a t-test is performed for each hexamer, between the "*ts*" and "*nts*" groups. Ranking the corresponding t-test P-values yields those hexamers that are most different distributionally between the positive and negative training samples. The top 1000 of these hexamers are chosen for further analysis. This step is only necessary to reduce the computational complexity of the overall procedure—computing the DI between each of the 4096 hexamers and the class label is relatively expensive.

(5) For the top $K = 1000$ hexamers which are most significantly different between the positive and negative training examples, $I(X_k^N \to Y^N)$ and $I(X_k^N; Y^N)$ reveal the degree of association for each of the $k \in (1, 2, \ldots, K)$ hexamers. The entropy terms in the directed information and mutual information expressions are found using a higher-order entropy estimator. Using the procedure of Section 7, the raw DI values are converted into their normalized versions. Since the goal is to maximize $I(X_k \to Y)$, we can rank the DI values in descending order.

(6) The significance of the DI estimate is obtained based on the bootstrapping methodology. For every hexamer, a $P = 0.05$ significance with respect to its bootstrapped null distribution yields potentially discriminative hexamers between the two classes. The Benjamini-Hochberg procedure is used for multiple-testing correction. Ranking the significant hexamers by decreasing DI value yields features that can be used for classifier (SVM) training.

(7) Train the support vector machine (SVM) classifier on the top d features from the ranked DI list(s). For comparison with the MI-based technique, we use the hexamers which have the top d (normalized) MI values. The accuracy of the trained classifier is plotted as a function of the number of features (d), after ten-fold cross-validation. As we gradually consider higher d, we move down the ranked list. In the plots below, the misclassification fraction is reported instead. A fraction of 0.1 corresponds to 10% misclassification.

Note. An important point concerns the training of the SVM classifier with the top d features selected using DI or MI (step (7) above). Since the feature selection step is decoupled from the classification step, it is preferred that the top d motifs are consistently ranked high among multiple draws of the data, so as to warrant their inclusion in the classifier. However, this does not yield expected results on this data set. Briefly, a kendall rank correlation coefficient [34] was computed between the rankings of the motifs between multiple data draws (by sampling a subset of the entire dataset), for both MI- and DI-based feature-selection. It is observed that this coefficient is very low in both MI and DI, indicating a highly variable ranking. This is likely due to the high variability in data distribution across these multiple draws (due to limited number of data points), as well as the sensitivity of the data-dependent entropy estimation procedure to the range of the samples in the draw. To circumvent this problem of inconsistency in rank of motifs, a *median* DI/MI value is computed across these various draws and the top d features based on the median DI/MI value across these draws are picked for SVM training [20].

11. RESULTS

11.1. Tissue specific promoters

We use DI to find hexamers that discriminate brain-specific and heart-specific expression from neutral sequences. The negative training sets are sequences that are not brain or heart-specific, respectively. Results using the MI and DI methods are given below (see Figures 5 and 7). The plots indicate the SVM cross-validated misclassification accuracy (ideally 0) for the data as the number of features using the metric (DI or MI) is gradually increased. We can see that for any given classification accuracy, the number of features using DI is less than the corresponding number of features using MI. This translates into a lower misclassification rate for DI-based feature selection. We also observe that as the number of features d is increased, the performance of MI is the same as DI. This is expected since, as we gather more features using MI or DI, the differences in MI versus DI ranking are compensated.

An important point needs to be clarified here. There is a possibility of sequence composition bias in the tissue-specific and neutral sequences used during training. This has been reported in recent work [15]. To avoid detecting GC rich sequences as hexamer features, it is necessary to confirm that there is no significant GC-composition bias between the specific and neutral sets in each of the case studies. This is demonstrated in Figures 4, 6, and 8. In each case, it is observed that the mean GC-composition is almost same for the specific versus neutral set. However, in such studies, it is necessary to select for sequences that do not exhibit such bias. In Figures 6 and 8, even the distribution of GC-composition is similar among the samples. For Figure 4, even though the distributions are slightly different, the box plots indicate similarity in mean GC-content.

Next, some of the motifs that discriminate between tissue-specific and nonspecific categories for the brain promoter, heart promoter, and brain enhancer cases, respectively, are listed in Table 2. Additionally, if the genes encoding for these TFs are expressed in the corresponding tissue [35], a (*) sign is appended. In some cases, the hexamer motifs match the consensus sequences of known transcription factors (TFs). This suggests a potential role for that particular TF in regulating expression of tissue-specific genes. This matching of hexamer motifs with TFBS consensus sites is done using the MAPPER engine (http://bio.chip.org/mapper). It is to be noted that a hexamer-TFBS match does not necessarily imply the functional role of the TF in the corresponding tissue (brain or heart). However, such information would be useful to guide focused experiments to confirm their role in vivo (using techniques such as chromatin immunoprecipitation).

As is clear from the above results, there are several other motifs which are novel or correspond to nonconsensus motifs of known transcription factors. Hence, each of the identified hexamers merit experimental investigation. Also, though we identify as many as 200 hexamers in this work (please see Supplementary Material available online at doi: 10.1155/2007/13853), we have reported only a few due to space constraints.

In the context of the heart-specific genes, we consider the cardiac troponin gene (*cTNT*, ENSEMBL: ENSG00000118194), which is present in the heart promoter set. An examination of the high DI motifs for the heart-specific set yields motifs with the GATA consensus site, as well as matches with the MEF2 transcription factor. It has been established earlier that GATA-4, MEF2 are indeed

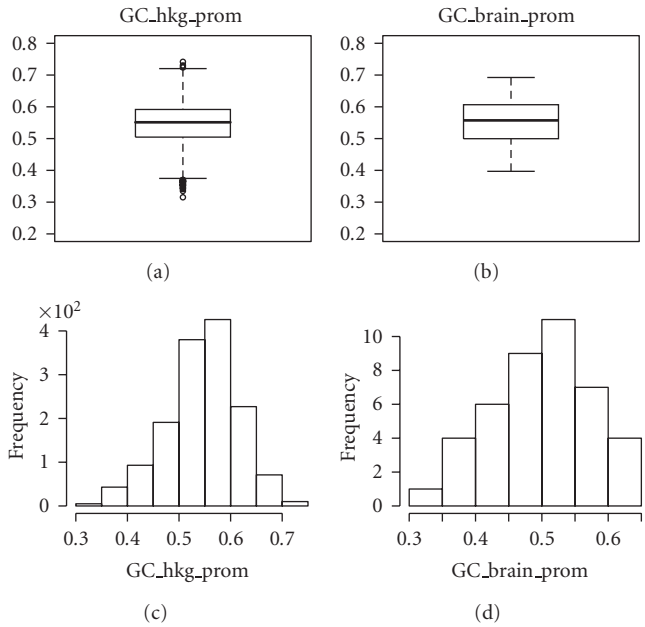

Figure 4: GC sequence composition for brain-specific promoters and housekeeping (hkg) promoters.

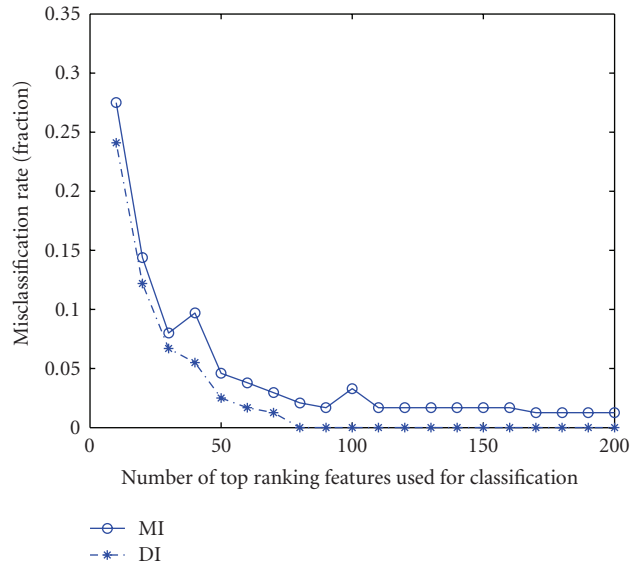

Figure 5: Misclassification accuracy for the MI versus DI case (brain promoter set). Accuracy of classification is ~0.9, that is, 93%.

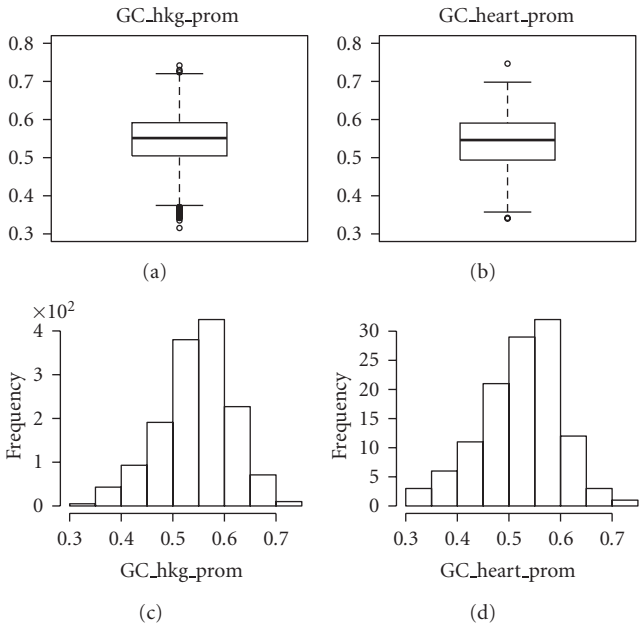

Figure 6: GC sequence composition for heart-specific promoters and housekeeping (hkg) promoters.

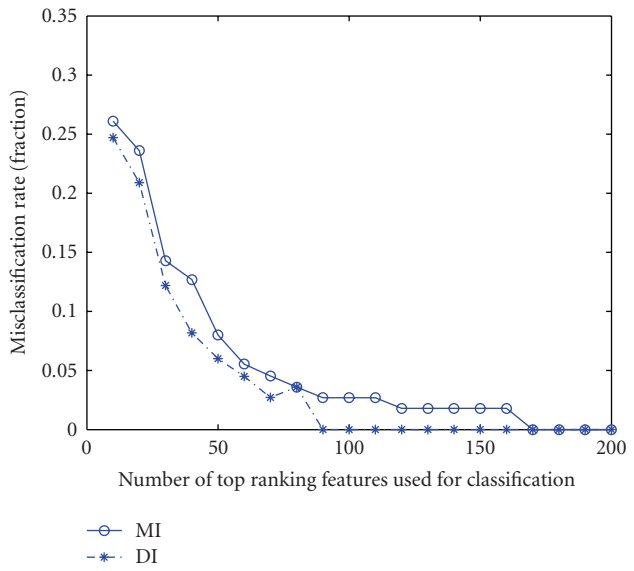

Figure 7: Misclassification accuracy for the MI versus DI case (heart promoter set).

involved in transcriptional activation of this gene [36] and the results have been confirmed by ChIP [37].

11.2. Enhancer DB

Additionally, all the brain-specific regulatory elements profiled in the mouse Enhancer Browser database (http://enhancer.lbl.gov) are examined for discriminating motifs. Figure 8 shows that the two classes have similar GC-composition. Again, the plot of misclassification accuracy

Table 2: Comparison of high ranking motifs (by DI) across different data sets. The (*) sign indicates tissue-specific expression of the corresponding TF gene.

Brain promoters	Heart promoters	Brain enhancers
Ahr-ARNT (*)	Pax2	HNF-4 (*)
Tcf11-MafG (*)	Tcf11-MafG (*)	Nkx2
c-ETS (*)	XBP1 (*)	AML1
FREAC-4	Sox-17 (*)	c-ETS (*)
T3R-alpha1	FREAC-4	Elk1 (*)
	GATA(*)	

versus number of features in the MI and DI scenarios reveal the superior performance of the DI-based hexamer selection compared to MI (see Figure 9).

In this case, the enhancer sequences are ultraconserved, thus obtained after alignment across multiple species. The examination of these sequences identified motifs that are potentially selected for regulatory function across evolutionary distances. Using alignment as a prefiltering strategy helps remove bias conferred by sequence elements that arise via random mutation but might be over-represented. This is permitted in programs like Toucan [12] and rVISTA (http://rvista.dcode.org).

As in the previous case, some of the top ranking motifs from this dataset are also shown in Table 2. The (*) signed TFs indicate that some of these discovered motifs indeed have documented high expression in the brain. The occurrence of such tissue-specific transcription factor motifs in these regulatory elements gives credence to the discovered motifs. For example, *ELK-1* is involved in neuronal differentiation [38]. Also, some motifs matching consensus sites of TEF1 and ETS1 are common to the brain-enhancer and brain-promoter set. Though this is interesting, an experiment to confirm the enrichment of such transcription factors in the population of brain-specific regulatory sequences is necessary.

11.3. Quantifying sequence-based TF influence

A very interesting question emerges from the above presented results. What if one is interested in a motif that is not present in the above ranked hexamer list for a particular tissue-specific set? As an example, consider the case for *MyoD*, a transcription factor which is expressed in muscle and has an activity in heart-specific genes too [39]. In fact, a variant of its consensus motif CATTTG is indeed in the top ranking hexamer list. The DI-based framework further permits investigation of the directional association of the canonical *MyoD* motif (CACCTG) for the discrimination of heart-specific genes versus housekeeping genes. This is shown in Figure 10. As is observed, *MyoD* has a significant directional influence on the heart-specific versus neutral sequence class label. This, in conjunction with the expression level characteristics of *MyoD*, indicates that the motif CACCTG is potentially relevant to make the distinction between heart-specific and neutral sequences.

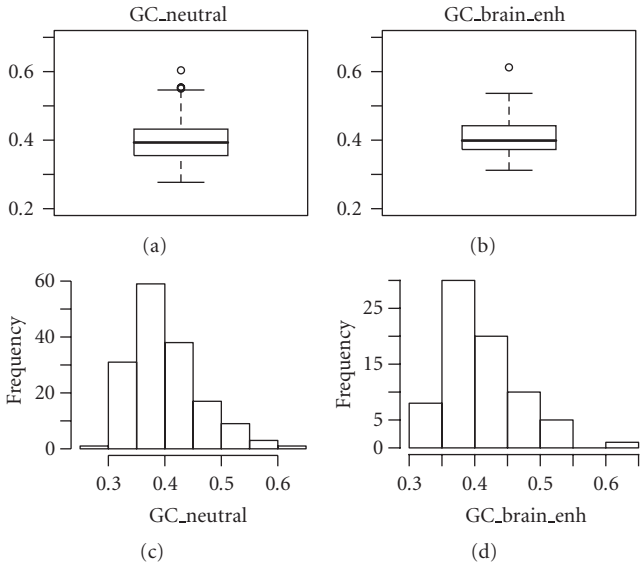

Figure 8: GC sequence composition for brain-specific enhancers and neutral noncoding regions.

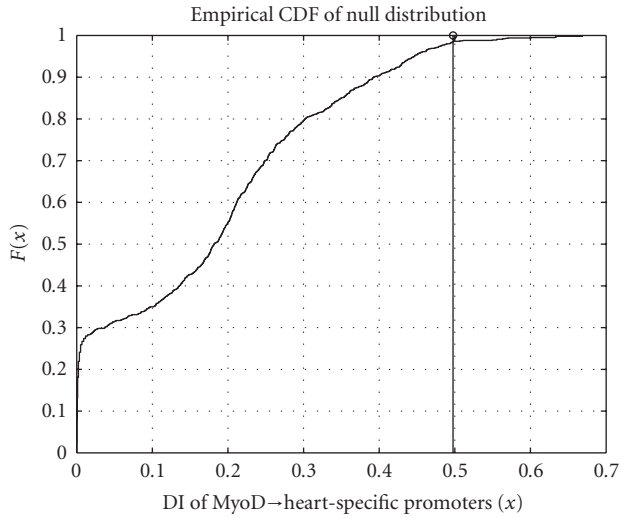

Figure 10: Cumulative distribution function for bootstrapped $I(MyoD$ motif: CACCTG$\rightarrow Y)$; Y is the class label (heart-specific versus housekeeping). True $\hat{I}(\text{CACCTG}\rightarrow Y) = 0.4977$.

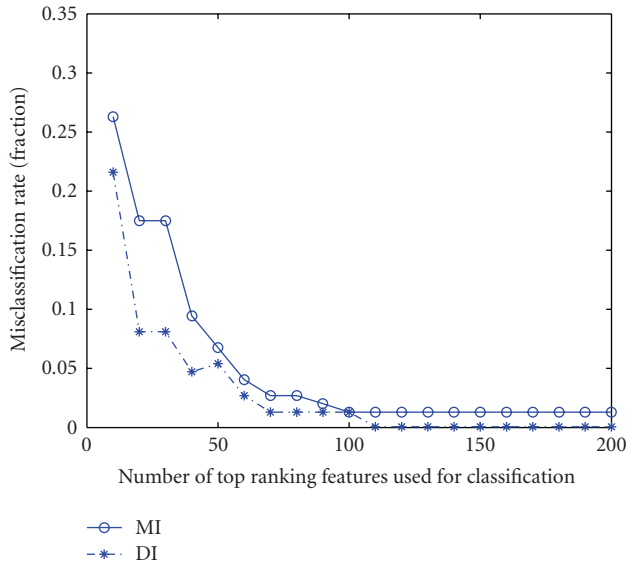

Figure 9: Misclassification accuracy for the MI versus DI case (brain enhancer set).

Another theme picks up on something quite traditionally done in bioinformatics research-finding key TF regulators underlying tissue-specific expression. Two major questions emerge from this theme.

(1) Which putative regulatory TFs underlie the tissue-specific expression of a group of genes?
(2) For the TFs found using tools like TOUCAN [12], can we examine the degree of influence that the particular TF motif has in directing tissue-specific expression?

To address the *first* question, we examine the TFs revealed by DI/MI motif selection and compare these to the TFs discovered from TOUCAN [12], underlying the expression of genes expressed on day *e*14.5 in the degenerating mesonephros and nephric duct (TS22). This set has about 43 genes (including *Gata2*). These genes are available in the Supplementary Material.

Using TOUCAN, the set of module TFs is combinations of the following TFs: *E47*, *HNF3B*, *HNF1*, *RREB1*, *HFH3*, *CREBP1*, *VMYB*, *GFI1*. These were obtained by aligning the promoters of these 43 genes (−2000 bp upstream to +200 bp from the TSS), and looking for over-represented TF motifs based on the TRANSFAC/JASPAR databases. Using the DI-based motif selection, a set of 200 hexamers are found that discriminate these 43 gene promoter sequences from the background housekeeping promoter set. They map to the consensus sites of several known TFs, such as (identified from http://bio.chip.org/mapper) *Nkx*, *Max1*, *c-ETS*, *FREAC4*, *Ahr-ARNT*, *CREBP2*, *E2F*, *HNF3A/B*, *NFATc*, *Pax2*, *LEF1*, *Max1*, *SP1*, *Tef1*, *Tcf11-MafG*; many of which are expressed in the developing kidney (http://www.expasy.org). Moreover, we observe that the TFs that are common between the TOUCAN results and the DI-based approach: *FREAC4*, *Max1*, *HNF3a/b*, *HNF1*, *SP1*, *CREBP*, *RREB1*, *HFH3,* are mostly kidney-specific. Thus, we believe that this observation makes a case for finding all (possibly degenerate) TF motif searches from TRANSFAC, and filtering them based on tissue-specific expression subsequently. Such a strategy yields several more TF candidates for testing and validation of biological function.

For the *second* question, we examine the following scenario. The *Gata3* gene is observed to be expressed in the developing ureteric bud (UB) during kidney development. To find UB specific TF regulators, conserved TF modules can be examined in the promoters of UB-specific genes. These experimentally annotated UB-specific genes are obtained from the Mouse Genome Informatics database at http://www.informatics.jax.org. Several programs are used for such analysis, like Genomatix [11] or Toucan [12]. Using

Toucan, the promoters of the various UB specific genes are aligned to discover related modules. The top-ranking module in Toucan contains *AHR-ARNT*, *Hox13*, *Pax2*, *Tal1alpha-E47*, *Oct1*. Again, the power of these motifs to discriminate UB-specific and nonspecific genes, based on DI, can be investigated.

For this purpose, we check if the *Pax2* binding motif (GTTCC [40]) indeed induces kidney specific expression by looking for the strength of DI between the GTTCC motif and the class label (+1) indicating UB expression (see Figure 11). This once again adds to computational evidence for the true role of *Pax2* in directing ureteric bud specific expression [40]. The main implication here is that from sequence data, there is strong evidence for the *Pax2* motif being a useful feature for UB-specific genes. This is especially relevant given the documented role of *Pax2* (see [41]) directing ureteric-bud expression of the *Gata3* gene, one of the key modulators of kidney morphogenesis. Both the *MyoD* and *Pax2* studies indicate the relevance of principled data integration using expression [35, 42] and sequence modalities.

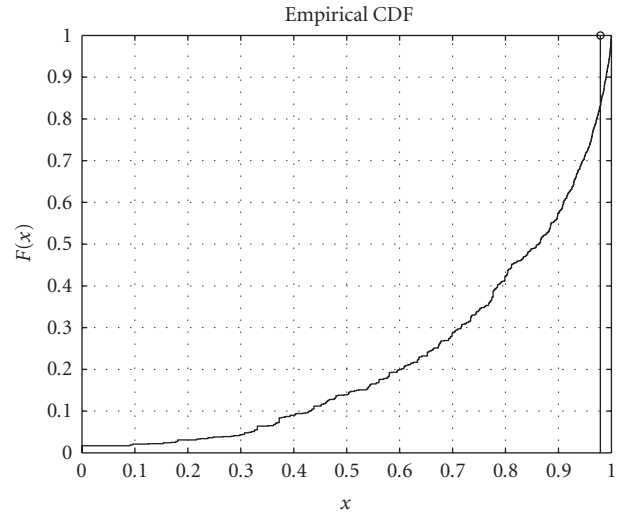

Figure 11: Cumulative distribution function for bootstrapped $I(Pax2\text{ motif: GTTCC}\rightarrow Y)$; Y is the class label (UB/non-UB). True $\widehat{I}(\text{GTTCC}\rightarrow Y) = 0.9792$.

11.4. Observations

With regard to the feature selection and classification results, in both studies (enhancers and promoters), we observe that about 100 hexamers are enough to discriminate the tissue-specific from the neutral sequences. Furthermore, some sequence features of these motifs at the promoter/enhancer emerge.

(i) There is higher sequence variability at the promoter since it has to act in concert with LREs of different tissue types during gene regulation.

(ii) Since the enhancer/LRE acts with the promoter to confer expression in only one tissue type, these sequences are more specific and hence their mining identifies motifs that are probably more indicative of tissue-specific expression.

We however, reiterate that the enhancer dataset that we study uses the *hsp68-lacz* as the promoter driven by the ultraconserved elements. Hence there is no promoter specificity in this context. Though this is a disadvantage and might not reveal all key motifs, it is the best that can be done in the absence of any other comprehensive repository.

The second aspect of the presented results highlights two important points. Firstly, the identified motifs have a strong predictive value as suggested by the cross-validation results as well as Table 2. Moreover, DI provides a principled methodology to investigate any given motif for tissue-specificity as well as for identifying expression-level relationships between the TFs and their target genes, (Section 11.3).

12. CONCLUSIONS

In this work, a framework for the identification of hexamer motifs to discriminate between two kinds of sequences (tissue-specific promoters or regulatory elements versus nonspecific elements) is presented. For this feature selection problem, a new metric—the "directed information" (DI)—is proposed. In conjunction with a support vector machine classifier, this method was shown to outperform the state-of-the-art method employing undirected mutual information. We also find that only a subset of the discriminating motifs correlate with known transcription factor motifs and hence the other motifs might be potentially related to non-consensus TF binding or underlying epigenetic phenomena governing tissue-specific gene expression. The superior performance of the directed-information-based variable selection suggests its utility to more general learning problems. As per the initial motivation, the discovery of these motifs can aid in the prospective discovery of other tissue-specific regulatory regions.

We have also examined the applicability of DI to prospectively resolve the functional role of any TF motif in a biological process, integrating other sources (literature, expression data, module searches).

13. FUTURE WORK

Several opportunities for future work exist within this proposed framework. Multiple sequence alignment of promoter/regulatory sequences across species would be a useful preprocessing step to reduce false detection of discriminatory motifs. The hexamers can also be identified based on other metrics exploiting distributional divergence between the samples of the "+1" and "−1" classes. Furthermore, there is a need for consistent high-dimensional entropy estimators within the small sample regime. A very interesting direction of potential interest is the formulation of a stepwise hexamer selection algorithm, using the directed information for maximal relevance selection and mutual information for minimizing between-hexamer redundancy [18]. This analysis is beyond the scope of this work but an implementation is available from the authors for further investigation. (The

source code of the analysis tools in R 2.0 and MATLAB 6.1 is available on request).

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the NIH under Award 5R01-GM028896-21 for J. D. Engel. They would like to thank Professor Sandeep Pradhan and Mr. Ramji Venkataramanan for useful discussions on directed information. They are extremely grateful to Professor Erik Learned-Miller and Dr. Damian Fermin for sharing their code for high-dimensional entropy estimation and ENSEMBL sequence extraction, respectively. They also thank the anonymous reviewers and the corresponding editor for helping them improve the quality of the manuscript through insightful comments and suggestions. The material in this paper was presented in part at the IEEE Statistical Signal Processing Workshop 2007 (SSP07).

REFERENCES

[1] K. D. MacIsaac and E. Fraenkel, "Practical strategies for discovering regulatory DNA sequence motifs," *PLoS Computational Biology*, vol. 2, no. 4, p. e36, 2006.

[2] G. Kreiman, "Identification of sparsely distributed clusters of *cis*-regulatory elements in sets of co-expressed genes," *Nucleic Acids Research*, vol. 32, no. 9, pp. 2889–2900, 2004.

[3] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic DNA," *Journal of Molecular Biology*, vol. 268, no. 1, pp. 78–94, 1997.

[4] Q. Li, G. Barkess, and H. Qian, "Chromatin looping and the probability of transcription," *Trends in Genetics*, vol. 22, no. 4, pp. 197–202, 2006.

[5] D. A. Kleinjan and V. van Heyningen, "Long-range control of gene expression: emerging mechanisms and disruption in disease," *The American Journal of Human Genetics*, vol. 76, no. 1, pp. 8–32, 2005.

[6] L. A. Pennacchio, G. G. Loots, M. A. Nobrega, and I. Ovcharenko, "Predicting tissue-specific enhancers in the human genome," *Genome Research*, vol. 17, no. 2, pp. 201–211, 2007.

[7] D. C. King, J. Taylor, L. Elnitski, F. Chiaromonte, W. Miller, and R. C. Hardison, "Evaluation of regulatory potential and conservation scores for detecting *cis*-regulatory modules in aligned mammalian genome sequences," *Genome Research*, vol. 15, no. 8, pp. 1051–1060, 2005.

[8] L. A. Pennacchio, N. Ahituv, A. M. Moses, et al., "In vivo enhancer analysis of human conserved non-coding sequences," *Nature*, vol. 444, no. 7118, pp. 499–502, 2006.

[9] K. Kadota, J. Ye, Y. Nakai, T. Terada, and K. Shimizu, "ROKU: a novel method for indentification of tissue-specific genes," *BMC Bioinformatics*, vol. 7, pp. 294, 2006.

[10] J. Schug, W.-P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and C. J. Stoeckert Jr., "Promoter features related to tissue specificity as measured by Shannon entropy," *Genome biology*, vol. 6, no. 4, p. R33, 2005.

[11] T. Werner, "Regulatory networks: linking microarray data to systems biology," *Mechanisms of Ageing and Development*, vol. 128, no. 1, pp. 168–172, 2007.

[12] S. Aerts, P. Van Loo, G. Thijs, et al., "TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis," *Nucleic Acids Research*, vol. 33, (Web Server Issue), pp. W393–W396, 2005.

[13] B. Y. Chan and D. Kibler, "Using hexamers to predict *cis*-regulatory motifs in Drosophila," *BMC Bioinformatics*, vol. 6, p. 262, 2005.

[14] G. B. Hutchinson, "The prediction of vertebrate promoter regions using differential hexamer frequency analysis," *Computer Applications in the Biosciences*, vol. 12, no. 5, pp. 391–398, 1996.

[15] P. Sumazin, G. Chen, N. Hata, A. D. Smith, T. Zhang, and M. Q. Zhang, "DWE: discriminating word enumerator," *Bioinformatics*, vol. 21, no. 1, pp. 31–38, 2005.

[16] G. Lakshmanan, K. H. Lieuw, K.-C. Lim, et al., "Localization of distant urogenital system-, central nervous system-, and endocardium-specific transcriptional regulatory elements in the GATA-3 locus," *Molecular and Cellular Biology*, vol. 19, no. 2, pp. 1558–1568, 1999.

[17] M. Khandekar, N. Suzuki, J. Lewton, M. Yamamoto, and J. D. Engel, "Multiple, distant *Gata2* enhancers specify temporally and tissue-specific patterning in the developing urogenital system," *Molecular and Cellular Biology*, vol. 24, no. 23, pp. 10263–10276, 2004.

[18] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[19] Proceedings of NIPS 2006 Workshop on Causality Feature Selection, http://research.ihost.com/cws2006/.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[21] H. Marko, "The bidirectional communication theory—a generalization of information theory," *IEEE Transactions on Communications*, vol. COM-21, no. 12, pp. 1345–1351, 1973.

[22] J. Massey, "Causality, feedback and directed information," in *Proceedings of the International Symposium on Information Theory and Its Applications (ISITA '90)*, pp. 303–305, Waikiki, Hawaii, USA, November 1990.

[23] R. Venkataramanan and S. S. Pradhan, "Source coding with feed-forward: rate-distortion theorems and error exponents for a general source," *IEEE Transactions on Information Theory*, vol. 53, no. 6, pp. 2154–2179, 2007.

[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, USA, 1991.

[25] E. G. Miller, "A new class of entropy estimators for multidimensional densities," in *Proceedings of the IEEE International Conference on Accoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 3, pp. 297–300, Hong Kong, April 2003.

[26] R. M. Willett and R. D. Nowak, "Complexity-regularized multiresolution density estimation," in *Proceedings of the International Symposium on Information Theory (ISIT '04)*, pp. 303–305, Chicago, Ill, USA, June-July 2004.

[27] I. Nemenman, F. Shafee, and W. Bialek, "Entropy and inference, revisited," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds., MIT Press, Cambridge, Mass, USA, 2002.

[28] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[29] H. Joe, "Relative entropy measures of multivariate dependence," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 157–164, 1989.

[30] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability, Chapman & Hall/CRC, Boca Raton, Fla, USA, 1994.

[31] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, Springer Series in Statistics, Springer, New York, NY, USA, 1997.

[32] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[33] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, NY, USA, 2001.

[34] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[35] NCBI Pubmed URL, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi.

[36] A. M. Murphy, W. R. Thompson, L. F. Peng, and L. Jones II, "Regulation of the rat cardiac troponin I gene by the transcription factor GATA-4," *Biochemical Journal*, vol. 322, part 2, pp. 393–401, 1997.

[37] A. Azakie, J. R. Fineman, and Y. He, "Myocardial transcription factors are modulated during pathologic cardiac hypertrophy in vivo," *The Journal of Thoracic and Cardiovascular Surgery*, vol. 132, no. 6, pp. 1262–1271.e4, 2006.

[38] P. Vanhoutte, J. L. Nissen, B. Brugg, et al., "Opposing roles of Elk-1 and its brain-specific usoform, short Elk-1, in nerve growth factor-induced PC12 differentiation," *Journal of Biological Chemistry*, vol. 276, no. 7, pp. 5189–5196, 2001.

[39] E. N. Olson, "Regulation of muscle transcription by the MyoD family: the heart of the matter," *Circulation Research*, vol. 72, no. 1, pp. 1–6, 1993.

[40] G. R. Dressler and E. C. Douglass, "Pax-2 is a DNA-binding protein expressed in embryonic kidney and Wilms tumor," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 4, pp. 1179–1183, 1992.

[41] D. Grote, A. Souabni, M. Busslinger, and M. Bouchard, "Pax2/8-regulated Gata3 expression is necessary for morphogenesis and guidance of the nephric duct in the developing kidney," *Development*, vol. 133, no. 1, pp. 53–61, 2006.

[42] A. Rao, A. O. Hero, D. J. States, and J. D. Engel, "Inference of biologically relevant gene influence networks using the directed information criterion," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, vol. 2, pp. 1028–1031, Toulouse, France, May 2006.