# DIRECTED-INFORMATION BASED FEATURE-SELECTION FOR TISSUE-SPECIFIC SEQUENCES

*Arvind Rao, Alfred O. Hero III, David J. States, James Douglas Engel*

Departments of EECS, Bioinformatics and Cell and Developmental Biology,
University of Michigan, Ann Arbor

## ABSTRACT

Motif discovery for the identification of functional regulatory elements underlying gene expression is a challenging problem. Sequence inspection often provides valuable clues to discovery of novel motifs (including transcription factor sites) with uncharacterized function in gene expression. Coupled with the complexity underlying tissue-specific gene expression, there are several motifs that are putatively responsible for gene 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 factors) and examine its application to several questions in current bioinformatics research. These motifs are seen to discriminate tissue-specific genomic regions from those that are not tissue-specific. We propose the use of directed information for such classification constrained feature selection, and then, use the selected features with a support vector machine (SVM) classifier to characterize the tissue-specificity of any sequence of interest. This analysis yields several novel interesting motifs that merit further experimental characterization. The last part of this paper presents a framework for exploring the relationship between such discriminatory transcription factor motifs, and the corresponding tissue-specificity, using both sequence and expression modalities.

***Index Terms***— Directed Information, transcriptional regulation, comparative genomics, tissue-specific genes.

## 1. INTRODUCTION

Gene expression via transcription involves the generation of messenger RNA from a DNA template and is a precursor to the production of protein via translation. Transcription involves the recruitment of transcription factor (TF) proteins at the gene's promoter as well as its long-range regulatory elements (such as enhancers). The computational prediction of regulatory elements genome-wide, is an interesting research question [5]. One approach to this question is to look for sequence motifs that are characteristic of tissue-specific gene expression. Thus, for any given gene, the motif underlying its expression in one tissue is potentially different from the motif conferring expression for that very gene in another tissue type. In this work, we consider the problem of motif discovery in the promoter and enhancer regions related to brain-specific gene expression.

## 2. ORGANIZATION

Initially, the set of tissue-specific regulatory regions (promoters and enhancers) are mined for their sequences. In section 3, the methodology for the data processing of these sequences into motif-sequence correspondence matrices is presented. In section 4, motif discovery is posed as a feature extraction problem, the utility of directed information in this framework is explained, and a metric for normalized directed information is proposed. Finally, a support vector machine (SVM) classifier is designed to discriminate the tissue-specific and non-specific sequences based on the hexamer motifs selected using DI (section 6). The paper concludes with results pertaining to the comparison of DI to MI for feature selection as well as demonstrates the utility of DI in more general bioinformatics problems pertaining to sequence selection.

## 3. DATA EXTRACTION AND PRE-PROCESSING

The Novartis foundation tissue-specificity atlas [*http://symatlas.gnf.org/*], has a compendium of genes and their corresponding tissues of expression. Genes have been profiled for expression in about twenty-five tissues, including adrenal gland, brain, dorsal root ganglion, spinal chord, testis, pancreas, liver etc. If a gene is expressed in less than three tissue types, it is annotated tissue-specific (*'ts'*), and if it is expressed in more than 22 tissue types, it is annotated to be non-specific (*'nts'*). Based on this assignment, we find a list of 45 genes that are tissue-specific as well as have brain expression. For these brain-specific genes, we extract their promoter sequences from the ENSEMBL database *http://www.ensembl.org/*], using sequence 2000bp upstream and 1000bp downstream up to the first exon

Funding: NIH 5R01-GM028896-21 (J.D.E).

relative to the transcriptional start site reported in ENSEMBL (release 37).

Before proceeding to motif selection, a matrix of motif-promoter correspondences is created. In this matrix, the counts of hexamer (six-nucleotide) motif occurrence in the *'ts'* and *'nts'* promoters is obtained using sequence parsing. The motif length of six is not overly restrictive, since it corresponds to the consensus binding site size of several annotated transcription factor motifs in the TRANSFAC/JASPAR databases. A Welch t-test is then performed between the relative counts of each hexamer in the two expression categories (*'ts'* and *'nts'*) and the top 1000 hexamers with $p - value \leq 10^{-6}$ are selected. This set of discriminating hexamers is designated ($\overrightarrow{\mathbf{H}} = H_1, H_2, \ldots, H_{1000}$). This procedure resulted in two hexamer-gene co-occurrence matrices, - one for the *'ts'* (or $+1$) class of dimension $N_{train,+1} \times 1000$ and the other for the *'nts'* (or $-1$) class - dimension $N_{train,-1} \times 1000$. Here $N_{train,+1}$ is the matrix of the 45 brain-specific genes. $N_{train,-1}$ is the set of *'nts'* that do not have brain-specific expression.

In the co-occurrence matrix, let $gc_{i,k}$ represent the absolute count of the $k^{th}$ hexamer, $k \in 1, 2, \ldots, M$ in the $i^{th}$ gene. Then, for each gene $g_i$, the quantile labeled matrix has $X_{i,k} = l$ if $gc_{i,[\frac{l-1}{K}M]} \leq gc_{i,k} < gc_{i,[\frac{l}{K}M]}$, $K = 4$. Matrices of dimension $N_{train,+1} \times 1001$, $N_{train,-1} \times 1001$ for the specific and non-specific 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 $(Y = -1/+1)$.

All the above steps, from promoter sequence extraction, parsing and quantization to obtain hexamer-promoter counts that are done for the brain-specific genes can be repeated for the brain-specific enhancers. This dataset is obtained from the Enhancer Browser [9], and contains 64 sequences that confer brain-specific expression in transgenic animals. Here, the 1000 hexamers discriminating brain-enhancers ($+1$ set) and a neutral set ($-1$), are designated $\overrightarrow{\mathbf{H'}} = H'_1, H'_2, \ldots, H'_{1000}$.

As an illustration, we show a representative matrix (Table 1).

| Ensembl Gene ID | AAAAAA | AAATAG | Class |
|---|---|---|---|
| ENSG00000155366 | 1 | 1 | +1 |
| ENSG000001780892 | 4 | 3 | +1 |
| ENSG00000189171 | 1 | 2 | -1 |
| ENSG00000168664 | 4 | 3 | -1 |
| ENSG00000160917 | 2 | 1 | -1 |
| ENSG00000176749 | 1 | 1 | -1 |
| ENSG00000006451 | 3 | 2 | +1 |

**Table 1**. The 'motif frequency matrix' for a set of gene-promoters. The first column is their ENSEMBL gene identifiers, the next 2 columns are hexamer quantile labels, and the last column is the corresponding gene's class label $(+1/-1)$.

## 4. DIRECTED INFORMATION AND FEATURE SELECTION

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 $j^{th}$ 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 [7]:

$$I(X_i^N \to Y^N) = \sum_{n=1}^{N} I(X_i^n; Y_n | Y^{n-1}) \qquad (1)$$

Using a stationarity assumption over a finite-length memory of the training samples, a correspondence with the setup in ([7], [12]) can be seen. As already known [1], the mutual information $I(X; Y) = H(X) - H(X|Y)$, where $H(X)$ and $H(X|Y)$ are the Shannon entropy of $X$ and the conditional entropy of $X$ given $Y$, respectively. With this definition of mutual information, the Directed Information simplifies to,

$$I(X^N \to Y^N) = \sum_{n=1}^{N} [H(X^n|Y^{n-1}) - H(X^n|Y^n)]$$

$$= \sum_{n=1}^{N} \{[H(X^n, Y^{n-1}) - H(Y^{n-1})] - [H(X^n, Y^n) - H(Y^n)]\} \qquad (2)$$

Using (2), 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 [8] based adaptive partitioning of the observation space can handle $N = 5/6$ without much complexity.

The relationship between MI and DI is given by [7],
DI: $I(X^N \to Y^N) = \sum_{i=1}^{N} I(X^i; Y_i | Y^{i-1})$.
MI: $I(X^N; Y^N) = \sum_{i=1}^{N} I(X^N; Y_i | Y^{i-1})$.

From [12], 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, we can see that the expression for DI is permutation-variant (i.e., the value of the DI is dependent on the ordering of random variables). Thus, what we find instead is the $I_p(X^N \to Y^N)$, a DI measure for a particular ordering of the $N$ random variables (r.vs). The DI value for our purpose, $I(X^N \to Y^N)$ is an average over all possible sample permutations given by, $I(X^N \to Y^N) = \frac{1}{N!} \sum_{p=1}^{N!} I_p(X^N \to Y^N)$. 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 ordering). As can be readily observed, this problem is combinatorially complex, and hence, a monte-carlo sampling strategy is used for computing $I(X^N \to Y^N)$.

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

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 ($X_k$) for which $I(X_k^N \rightarrow Y^N)$ is the highest. From the definition of DI, we know that $0 \leq I(X_k^N \rightarrow Y^N) \leq I(X_k^N; Y^N) < \infty$. 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 $\rho_{DI}$ should be able to map this large range ($[0, \infty]$) to $[0, 1]$. Following [3], an expression for the normalized DI is given by:
$\rho_{I(X^N \rightarrow Y^N)} = \sqrt{1 - e^{-2I(X^N \rightarrow Y^N)}}$.

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 series $X_i$ and $Y$). This is done using confidence intervals after permutation testing (section: 5).

## 5. BOOTSTRAPPED CONFIDENCE INTERVALS

In the absence of knowledge of the true distribution of the DI estimate, an approximate confidence interval for the DI estimate ($\hat{I}(X^N \rightarrow Y^N)$), is found using bootstrapping [2]. Density estimation is based on kernel smoothing over the bootstrapped samples [10].

The kernel density estimate for the bootstrapped DI (with $n = 1000$ samples), $Z \triangleq \hat{I}_B(X^N \rightarrow Y^N)$ becomes,
$\hat{f}_h(Z) = \frac{1}{nh} \sum_{i=1}^{n} \frac{3}{4}[1 - (\frac{z_i - z}{h})^2] I(\left|\frac{z_i - z}{h}\right| \leq 1)$ with $h \approx 2.67\hat{\sigma}_z$ and $n = 1000$. $\hat{I}_B(X^N \rightarrow Y^N)$ is obtained by finding the DI for each random permutation of the $X$, $Y$ series, and performing this permutation $B$ times. As is the 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 AMISE (asymptotic mean squared error) and favorable bias-variance tradeoff [10].

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

## 6. SVM CLASSIFICATION

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. A 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 $N_s (= N_{train,+1} + N_{train,-1})$ pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_{N_s}, y_{N_s})$, with $x_i \in \mathscr{R}^d$ and $y_i \in \{-1, +1\}$.

An SVM is a maximum margin hyperplane classifier in a non-linearly 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 [2].

## 7. SUMMARY OF OVERALL APPROACH

Here, the term ‘sequence’ can pertain to either brain-specific promoters or brain-enhancer sequences.

- Parse sequences to obtain hexamer counts for the *‘ts’* and *‘nts’* set. Use preprocessing to create matrices $N_{train,\underline{+}1} \times 1001$ corresponding to the sequences in the *‘ts’* and *‘nts’* sets.

- For the $J = 1000$ hexamers find $I(X_j \rightarrow Y)$ and $I(X_j \rightarrow Y)$ for each of the $j \in (1, 2, \ldots, J)$ hexamers. Since the goal is to maximize $I(X_j \rightarrow Y)$, we can rank the $\rho_{DI}$ values in descending order.

- Find hexamers whose $\rho_{DI}$ is 0.05 significant with respect to its bootstrapped null distribution (using kernel density estimation), and rank the hexamers by decreasing $\rho_{DI}$ value. The top $'d'$ hexamers in this ranked list can be used for classifier (SVM) training.

- Train the Support Vector Machine classifier (SVM) on the top $'d'$ features from the ranked DI list. For comparison with the MI based technique, we use the hexamers which have the top $'d'$ MI values. The accuracy of the trained classifier as a function of the number of features ($d$) is plotted, after cross-validation. As we gradually consider higher $'d'$, we move down the ranked list. In the results below, the misclassification fraction is plotted instead. A fraction of 0.1 corresponds to 10% misclassification.

## 8. RESULTS

DI is used to find discriminating hexamers that underlie brain-specific expression. The negative training sets are sequences that are not brain-specific. Results using the MI and DI methods are given above (Figs. 2 and 4). The plots indicate the

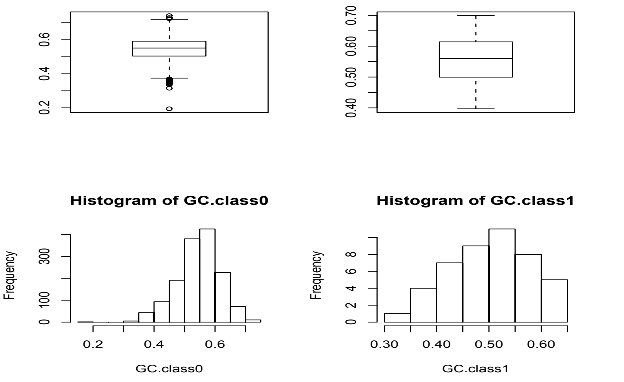

**Fig. 1**. GC sequence composition for brain-specific promoters and housekeeping promoters.

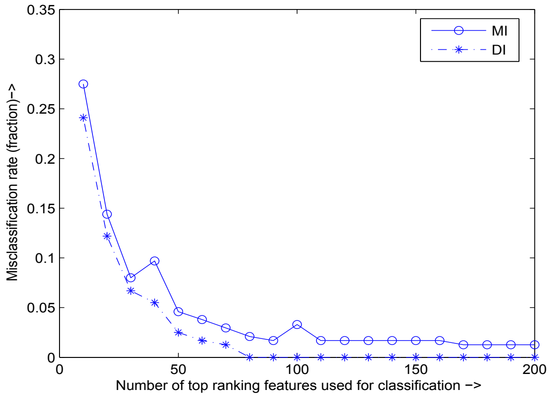

**Fig. 2**. Misclassification accuracy for the MI vs. DI case (brain promoter set). Accuracy of classification is $\sim 0.9$ i.e. 93%.

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 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 variations in MI/DI ranking are compensated. Several brain-specific motifs rank high in the DI-based ranking (e.g.: *c-ETS*, *Elk1*, *Ahr-ARNT*, *GTTCCA*).

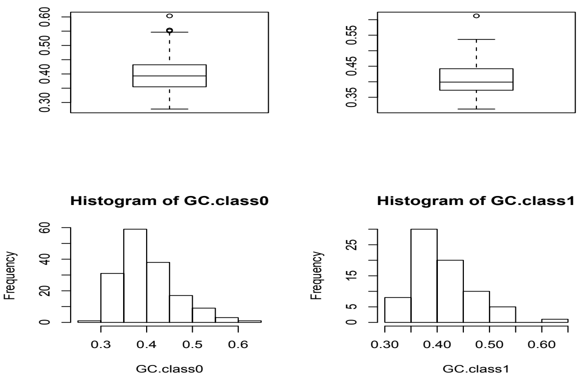

**Fig. 3**. GC sequence composition for brain-specific enhancers and neutral non-coding regions.

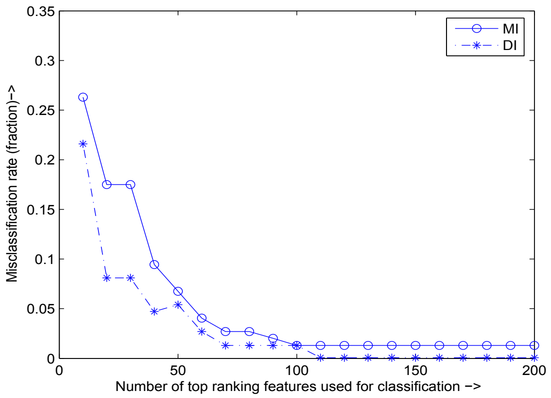

**Fig. 4**. Misclassification accuracy for the MI vs. DI case (brain enhancer set).

An important point needs to be clarified. Sequence composition bias is a confounding factor in the analysis of tissue-specific and non-specific sequences. It is thus possible that the motifs that are selected are just GC-rich because of the higher GC composition of tissue-specific sequences. To avoid this problem, the selected sequences are checked for GC-composition and the box plots for the composition across the samples as well as the distribution is given in Figs. 1 and 3. These plots show that the GC content of these sequences are distributionally similar between the 'ts' and 'nts' sequences (promoters and enhancers) - thereby avoiding bias.

Some of the top ranking motifs from this dataset are also shown in Table 2. As indicated by the (*) signed TFs, some of these discovered motifs indeed have documented high expression in the brain. For example, *ELK-1* is involved in neuronal differentiation [11]. 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.

From the results above, the following observations can be made:

| Brain promoters | Brain enhancers |
|---|---|
| Ahr-ARNT (*) | HNF-4 (*) |
| Tcf11-MafG (*) | Nkx |
| c-ETS (*) | AML1 |
| FREAC-4 | c-ETS (*) |
| T3R-alpha1 | Elk1 (*) |

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

- There is more sequence variability at the promoter since it has to act in concert with enhancers of different tissue types.
- 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.

### 8.1. 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 *Nurr1*, a transcription factor which is expressed in brain and has an important role in central nervous system (CNS) development [4]. In fact, a variant of its consensus motif - TCCAGA is indeed in the top ranking hexamer list. The DI based framework further permits investigation of the directional association of the *Nurr1* motif (TCCAGA) for the discrimination of brain-specific genes vs. housekeeping genes (Fig. 5). As is observed, *Nurr1* has a significant directional influence on the brain-specific vs. neutral sequence class label. This, in conjunction with the expression level characteristics of *Nurr1*, indicates that the motif TCCAGA is potentially relevant to make the distinction between brain-specific and neutral sequences.

## 9. CONCLUSIONS

In this work, we have presented a framework for the identification of hexamer motifs to discriminate between two kinds of sequences (tissue-specific promoters or regulatory elements vs non-specific elements). For this feature selection problem we proposed the utility of a new metric - the 'directed information' (DI). In conjunction with a support vector machine classifier, this method was shown to outperform the state-of-the-art methods employing undirected mutual information. We also find that only a subset of the discriminating motifs correlate with known transcription factor motifs and hence might be potentially related to underlying interesting phenomena governing tissue-specific expression. Finally, we demonstrate that DI can be used to find the discriminatory potential of any chosen motif (*Nurr1*, in this case) between a set of tissue-specific and non-specific sequences. The superior performance of the directed-information based variable selection suggests its utility to more general learning problems.

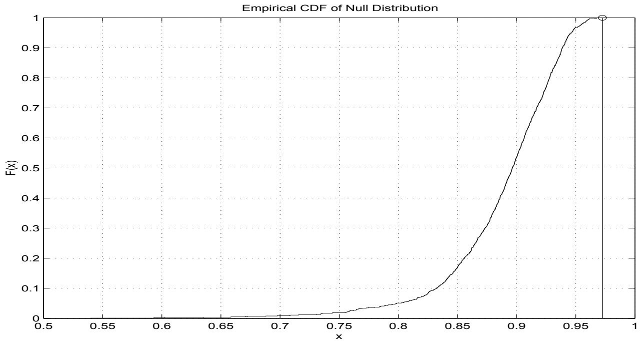

**Fig. 5**. Cumulative Distribution Function for bootstrapped $I(\mathit{Nurr1\ motif{:}TCCAGA} \rightarrow Y)$; $Y$ is the class label (Brain-specific vs. Housekeeping). True $\hat{I}(\mathit{TCCAGA} \rightarrow Y) = 0.9725$.

### 10. REFERENCES

[1] Cover TM, Thomas JA, Elements of Information Theory, *Wiley- Interscience*, 1991.

[2] Hastie T, Tibshirani R, The Elements of Statistical Learning , Springer 2002.

[3] H. Joe., "Relative entropy measures of multivariate dependence", *J. Am. Statist. Assoc.*, 84:157164, 1989.

[4] Law SW, Conneely OM, DeMayo FJ, O'Malley BW.,"Identification of a new brain-specific transcription factor, NURR1", *Mol Endocrinol.* 1992 Dec;6(12):2129-35.

[5] MacIsaac KD, Fraenkel E., "Practical strategies for discovering regulatory DNA sequence motifs", *PLoS Comput Biol.* 2006 Apr;2(4):e36.

[6] H. Marko, "The Bidirectional Communication Theory - A Generalization of Information Theory", *IEEE Transactions on Communications*, Vol. COM-21, pp. 1345-1351, 1973.

[7] J. Massey, "Causality, feedback and directed information," *Proc. 1990 Symp. Information Theory and Its Applications (ISITA-90)*, Waikiki, HI, Nov. 1990, pp. 303305.

[8] Learned-Miller E., "Hyperspacings and the estimation of information theoretic quantities", UMass Amherst Technical Report 04-104, 2004.

[9] Pennacchio, L. A., Ahituv, N., Moses, A., Prabhakar, S., Nobrega, M., Shoukry, M., Minovitsky, A., Dubchak, I., Holt, A., Lewis, K., Plazer-Frick, I., Akiyama, J., DeVal, S., Afzal, V., Black, B., Couronne, O., Eisen, M., Visel, A., and Rubin, E.M. 2006., "In vivo enhancer analysis of human conserved non-coding sequences", *Nature*, 444(7118):499-502.

[10] J. Ramsay, B. W. Silverman, Functional Data Analysis (Springer Series in Statistics), Springer 1997.

[11] Vanhoutte P, Nissen JL, Brugg B, Gaspera BD, Besson MJ, Hipskind RA, Caboche J., "Opposing roles of Elk-1 and its brain-specific isoform, short Elk-1, in nerve growth factor-induced PC12 differentiation", *J Biol Chem.* 2001 Feb 16;276(7):5189-96.

[12] Venkataramanan, R.; Pradhan, S. S.,"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, Jun. 2007.