African Journal of Biotechnology Vol. 10(64), pp. 13974-13981, 19 October, 2011
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB10.1527
ISSN 1684–5315 © 2011 Academic Journals

Full Length Research Paper

DNA regulatory motif selection based on support vector machine (SVM) and its application in microarray experiment of Kashin-Beck disease

XiaoMing Wu[1,2], JianQiang Du[1,2], Shuang Wang[1,2], Min Zhang[1,2], Xuanqi Wang[3] and Xiong Guo[1,2]*

[1]The Key Laboratory of Biomedical Information Engineering, Ministry of Education, School of Life Science and Technology, P. R. China.
[2]The Key Laboratory of Environment and Genes Related to Diseases, Ministry of Education, Medicine College, Xi'an Jiaotong University, Xi'an 710049, P. R. China.
[3]Department of Cardiology, the Fourth People's Hospital of Shaanxi Province, Xi'an 710043, P. R. China.

Accepted 8 April, 2011

Conserved DNA sequences are essential to investigate the regulation and expression of nearby genes. The conserved regions can interact with certain proteins and can potentially determine the transcription speed and amount of the corresponding mRNA in gene replication process. In this paper, motifs of co-expressed genes of microarray experiments were explored with discovery algorithms. Then a selection algorithm based on support vector machine (SVM) was applied to identify those motifs which mostly influenced gene expression. This method combined the advantages from both matrix based motif finding and functional motif selection. When applied to Kashin-Beck disease (KBD), this method identified 9 motifs, and revealed that some motifs may be related to the immune reactions. In addition, we suggested that the methods used could be applied to other microarray experiments to explore the underlying relationships between motif types and gene functions.

Key words: Support vector machine (SVM), microarray, motif discovery, gene regulation, Kashin-Beck disease.

INTRODUCTION

Kashin-Beck disease (KBD) is a chronic, endemic osteochondropathy affecting more than 2.5 million patients, and as much as 30 million people are prone to this disease in China only (Wang et al., 2009). The basic pathological features of the disease are degeneration and necrosis, mainly in growth plate cartilage and articular cartilage, which can result in growth retardation, secondary osteoarthrosis and disability in advanced stages. Clinically, the disease manifests as enlarged interphalangeal joints, shortened fingers, as well as restrained motion of joints. Since the pathogenesis of KBD is still unknown, the identification of differently expressed genes of microarray experiments and the investigation of basic mechanisms which led to the gene expression may shed light on the cause of the disease.

Understanding the relationship between DNA motifs in the upstream regions of genes is one of the most important aims for the current bioinformatical research. The importance of DNA-binding proteins in molecular functions such as transcription, replication, DNA repair and chromosome segregation highlight the significance of identifying the locations of their binding sites throughout the genome (Spyrou et al., 2009). It is well known that the gene expression is regulated by some transcription factors interacting with special binding sites of promoter on the DNA molecule near the gene transcription start site (TSS). However, different genes may have different sets of binding sites, so it is challenging to identify them efficiently. TRANSFAC, TRRD and JASPAR (Portales-Casamar et al., 2010) are three popular databases containing well known DNA motifs which can bind to different proteins. Some motifs in the databases were experimentally proved to be significant and can cause

*Corresponding author. E-mail: guox@mail.xjtu.edu.cn. Tel: +86-29-82655091. Fax: +86-29-82663454.

gene regulation, but most of which were attested from the alignments of experimental tested sequences. Many unmatched motif instances in genome are not included in the databases.

Computational method

In human promoters, the CAAT box, SP1 and TATA box are recognized by the constitutive transcription factors NF-Y, SP1 and TBP respectively, and are thought to be localized near the TSS (FitzGerald et al., 2004). But in many circumstances, researchers have little prior knowledge of the kind of other binding site that exists in each gene. Even if the DNA sequence was compared with entries in database, the part of sequence that can bind to protein is still unknown. Through motif discovery method on a set of DNA sequences derived from upstream regions of co-expressed genes, new motifs whose functions are related to the genes expression can be possibly discovered.

Discovery algorithm

Computational methods have been proven to be extremely effective in identifying TF-binding motifs, which are usually modeled by an A position weight matrix (PWM). Currently, many motif discovery methods have emerged. For example, Consensus, Gibbs sampler, W-AlignACE (Chen et al., 2008), MEME (Bailey and Elkan, 1994) and MDSCAN (Liu et al., 2002) are frequently used for sequence analysis.

Although motif could be discovered, further analysis is still needed. MotifExpress can make use of motif occurrence values, in combination with the gene express value, to select the most important motifs by using a regression method. It uses not only the sequence information but also the gene expression measurement (Zamdborg and Ma, 2009). Through database search service such as STAMP (Mahony and Benos, 2007), researchers could compare the canonic motif and putative motifs to investigate their functions.

In this study, 4 microarray experiments on KBD were conducted and the differentially expressed genes were selected by *t*-test. The over-represented sequence motifs were then discovered by Mdmodule (Conlon et al., 2003) on upstream region of these genes, and then, a support vector machine (SVM) based selection method was devised and was used to select the most important motifs and remove spurious motifs in up-regulated and down-regulated genes. Finally, the discovered motifs were summarized in detail.

MATERIALS AND METHODS

Overall strategy

There were 4 steps to achieve the goals in this study. Firstly, the microarray experiments were made on the blood samples of patients and the control individuals to obtain the expression value of each gene. Secondly, the increased and decreased mRNAs were selected by *t*-test, and the upstream sequences of each gene were retrieved from genome database thereafter to construct a sequence dataset. Thirdly, Mdmodule (Conlon et al., 2003), a fast and sensitive motif finding method, was used to generate a large set of motif enriched in the DNA sequences of these genes. In order to select the most special motifs relating to the gene expression types, a SVM classification method was used. The principle of the method is to calculate the precision of the SVM classifier on randomly selected subset of motifs. Finally, the functions of the selected motifs were further explored in detail. A flowchart of the process is shown in Figure 1.

Gene regulation model

Although gene is a long chain of deoxyribonucleic acid enchased in a huge bio-molecule of chromosome, the structure of a gene can be considered as many DNA motifs existing in the upstream region of TSS, with sequence segments containing coding exons. Gene is regulated by different kind of TFs binding to these regions and incurs succedent DNA formation alternation and gene replication. The amount of genes produced can be measured by microarray experiment, and the up-regulated genes and down-regulated genes can be identified by comparing the fluorescence intensity in the case and control samples. We can model the basal ideas of gene expression: denoting the gene as $G_i = (m_{i,1}, m_{i,2}, ...m_{i,k}; y_i)$, where, i is the genes index, k is the total number of motif type in the upstream region, $m_{i,k}$ represents the motif existence in the promoter region of the gene and y_i is the experiment result of microarray, which is either –1 or 1 (y_i = 1, up-regulated; y_i = -1, down-regulated). In this model, the level of gene expression is represented by the motif type in the promoter region. In fact, gene expression is not just determined by the motif type or the motif location, other signal molecular, even environment factors also play roles. But these factors were not included in this sample model. If some motifs are critical to a gene's expression and their influence force is high, our advanced model can take effect very well.

Motif identification

Generally, regulatory motifs are present in the promoter regions of gene; the DNA sequence of the specific gene in these regions can be downloaded from human genome database. All the upstream sequences of target genes were collected to construct a data set. Then, Mdmodule, a motif discovery method, was used to discover the candidate motifs. In the parameter, the motif length was set from 5 to 15. The value of a DNA sequence matching a motif was calculated by:

$$S_{mg} = \log_2[\sum_{x \in X_{mg}} \Pr(x \text{ from } \theta_m) / \Pr(x \text{ from } \theta_0)] \quad (1)$$

Where, θ_m is the probability matrix of motif m of width w and θ_0 is the third-order Markov model estimated from background DNA sequence. x_{wg} is the set of all w-mers in the upstream sequence of gene G (Liu et al., 2002).

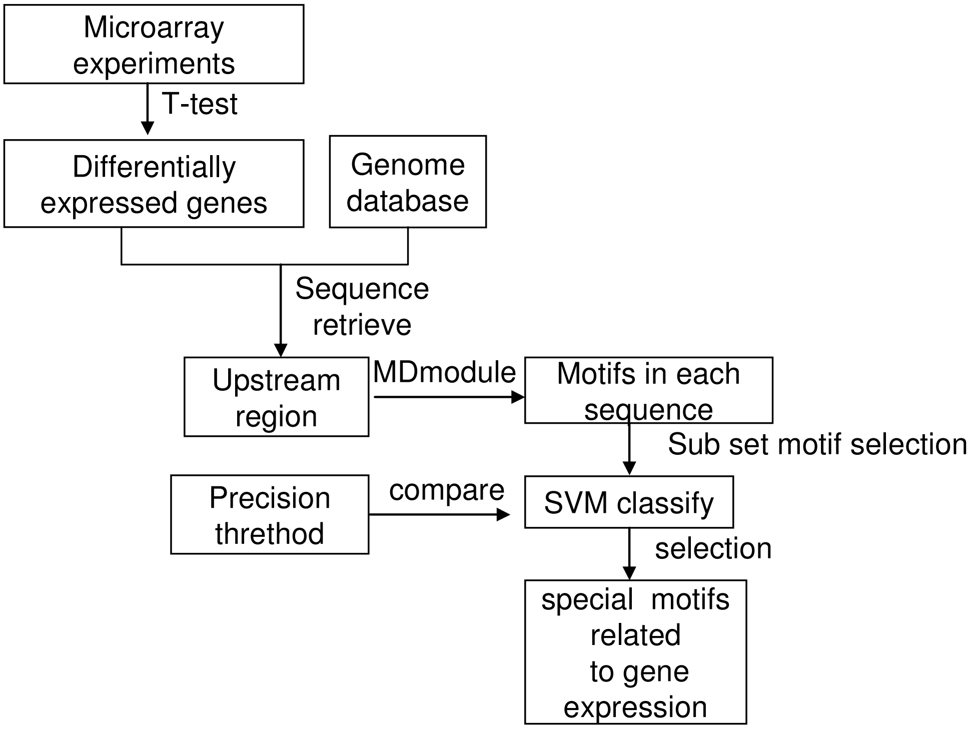


Figure1. Overall research strategy.

SVM selection of motifs

SVM is a general discriminative method to make classification. Through kernel function transformation, different input data can be handled. SVM has demonstrated high classification accuracy in molecular biology study such as protein family prediction (Jaakkola et al., 1999) and gene expression classification (Brown et al., 2000; Seok et al., 2010).

In this study, not all the motifs in the DNA sequences were functional motif or shared same level of importance, and there were many noisy signal motifs. Since each sequence may contain many motif types, it was a high-dimensional data. Through SVM, we can find out the most biological functional motifs. As a margin classifier, SVM draws an optimal hyperplane in a high-dimensional feature space, and maximizes the margin between data samples in two classes, therefore giving good generalization properties (He et al., 2006).

In this study, each motif had a matching score in every DNA sequence, and we denoted it as x_i. The regulation type was denoted as y_i, where $y_i = 1$ it means the gene was up-regulated, and $y_i = -1$ means the gene was down-regulated. So each gene can be designated as:

$$g_i = \{x_{i1}, x_{i2}, ..., x_{im}, y_i\}$$

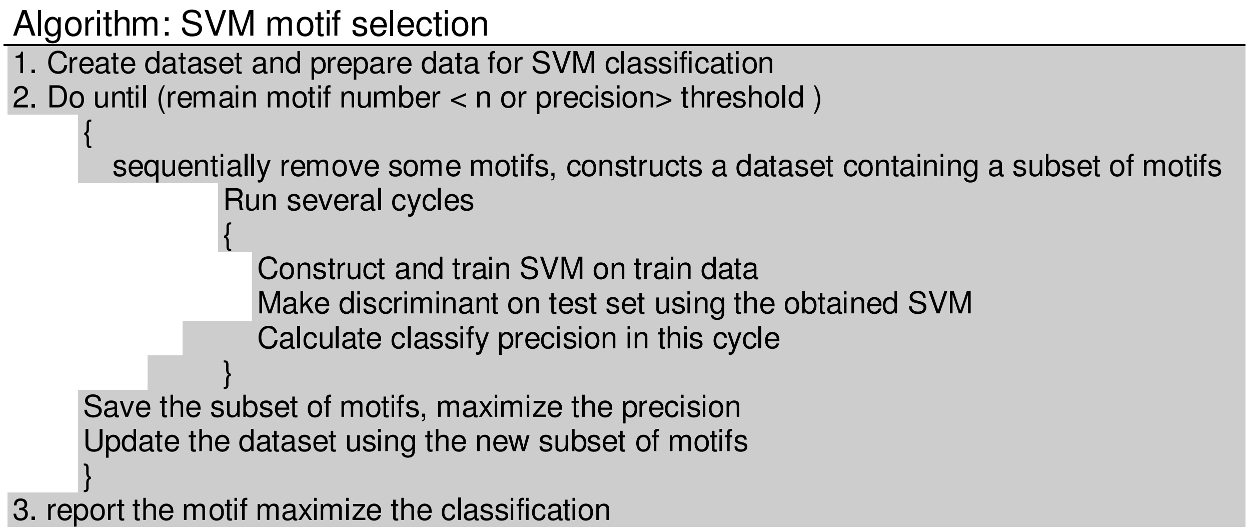
Here, $\vec{x} = x_1, x_2, ..., x_k$, $x_i = x_{i1}, x_{i2}, ..., x_{im}$ is a set of training examples, and $y = y_1, y_2, ..., y_k$ is the corresponding set of classifications. We considered that the motifs in the DNA sequence influenced the expression type of the gene, so we used SVM to select the motifs mostly related to the gene expression. The pseudocode for the algorithm is given in Figure 2.

By setting different threshold and parameter, we could run the earlier mentioned procedure many times; the motifs of particular interest were discovered finally.

RESULTS AND DISCUSSION

Microarray experiment

Controls and KBD cases were selected according to "National Clinical Criteria to diagnose KBD in China" (GB 16395-1996) (Gentleman et al., 2004). The investigation was approved by the Human Ethics Committee, Xi'an Jiaotong University. All patients and control set individuals approved the informed consents. The blood samples were collected from Xianyang, Shaanxi Province, China. 3 ml of peripheral blood were collected in each case of the 20 KBD patients, and then each 5 samples were randomly mixed to form 4 specimens. For comparison purpose, 3 ml peripheral blood of 12 healthy individuals of same area was also used to form 4 comparison specimens (Table 1). The microarray experiments were performed by Shaanxi Lifegen Co. Ltd., using 4 agilent human 1A oligo microarray chips. In each experiment, 21073 credibility gene expression values of controls and



```
Algorithm: SVM motif selection
1. Create dataset and prepare data for SVM classification
2. Do until (remain motif number < n or precision> threshold )
      {
         sequentially remove some motifs, constructs a dataset containing a subset of motifs
                     Run several cycles
                     {
                        Construct and train SVM on train data
                        Make discriminant on test set using the obtained SVM
                        Calculate classify precision in this cycle
                     }
      Save the subset of motifs, maximize the precision
      Update the dataset using the new subset of motifs
      }
3. report the motif maximize the classification
```

Figure 2. Algorithm of SVM motif selection.

Table 1. Comparison of the age and gender among four groups.

Group	KBD patient				Health control			
	n	**Average age (range)**	**Male**	**Female**	**n**	**Average age (range)**	**Male**	**Female**
1	5	47.20 (41-55)	2	3	3	45.00 (35-63)	1	2
2	5	50.40 (43-58)	2	3	3	48.67(39-68)	1	2
3	5	50.20 (40-67)	2	3	3	44.67(37-59)	1	2
4	5	46.60 (38-58)	3	2	3	41.00 (27-54)	0	3
Total	20	48.60 (38-67)	9	11	12	44.84 (27-68)	3	9

n is the number of blood samples and range represents the youngest and the oldest individuals.

cases were obtained, and the total values were 21073*2*4.

Differently expressed genes selection

Gene expression value normalization

The original data contain some system errors, which must be eliminated before gene selection process. In this study, locally weighted linear regression (LWR) method was used to correct such errors. Particularly in parameter setting, the default window size is 0.05 (5% of total data) and the order of the algorithm is 1. For comparison purpose, we made a scatter diagram of each gene chip.

Figure 3 illustrates the data points before and after normalization. An obvious data quality improvement can be seen. After normalization, the data points were distributed near the diagonal line; while before normalization, the data points departed from the diagonal line. Thus, the normalization makes the distribution more acceptable and more suitably to make further process of gene selection.

Differentially expressed genes selection by t-test

The t-test assesses whether the means of two groups are statistically different from each other. In this study, the expression values of each gene of the ase and control were compared. Each sample contained 4 values corresponding to 4 microarray experiments. The p values on the normalized data were calculated to find differently expressed genes. 168 genes were selected as differentially expressed at 0.01 significant level which were further selected to extract upstream sequences to construct the dataset.

Binding sites discovered from these genes

Dataset construction

After genes selection, their upstream regions related to TSS were extracted and a sequence dataset was constructed. Ensembl is a genomic database containing the basic gene information such as genomic position, gene function and gene sequences. In this study, the

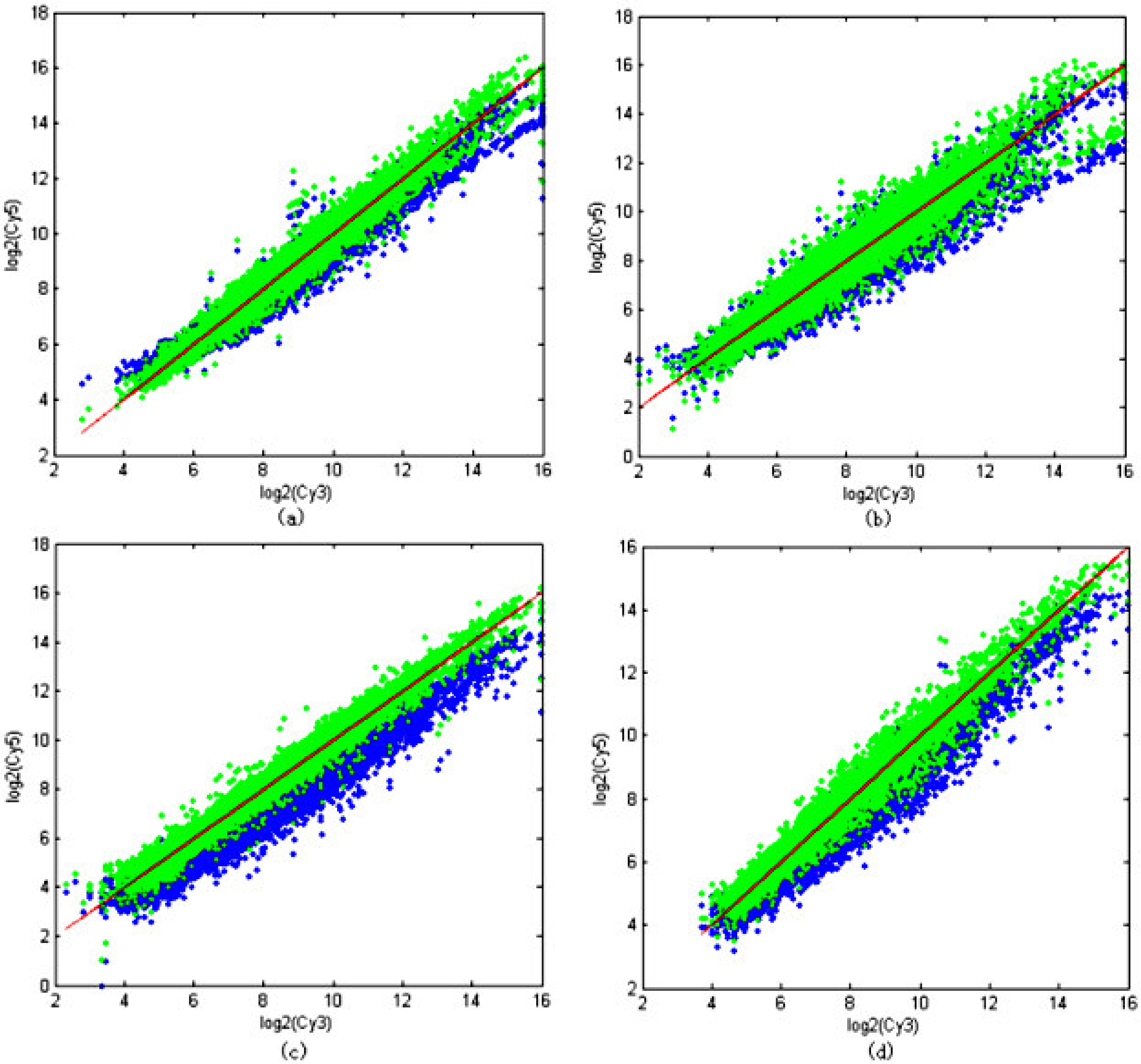


Figure 3. Before (blue) and after (green) locally weighted linear regression.

genomic position of each gene was located according to the Ensembl database (release 42 Dec 2006), then 2000 bp long upstream sequences of each gene were downloaded from web interface, and a sequences dataset was then constructed. Totally, 168 upstream sequences were included in the dataset. In the dataset, 86 genes were in the positive chain of the chromosome, while 82 genes were in the negative chain. In these genes, 88 of them were up-regulated and 80 were down-regulated.

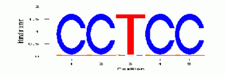
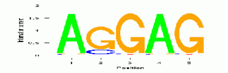
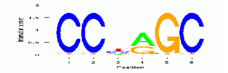
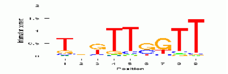
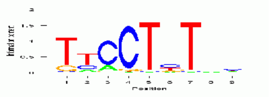
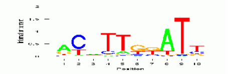
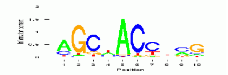
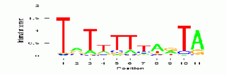
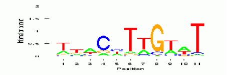
Motifs discovered

We used Mdmodule (Lu et al., 2008) to search the upstream sequences. It generated 186 candidate motifs for each width from 5 to 15 bases. The discovered motifs were saved in a PWM format. Also, every motif was scanned along each DNA sequence, and scores of motif related to each sequence were calculated by Mdmodule algorithm using default parameters. Each motif was transformed to a consensus sequence and a logo using Seqlogo program in bioconductor package (Gentleman et al., 2004).

Motif sites selected using SVM

Each motif has a value indicating its existence in each DNA sequence. As a result, many values would relate to the expression of one single gene. The dataset were divided into 2 classes: up-regulated genes and down-regulated genes, labeled as -1 or 1. The values of each

Table 2. The top 9 motifs with maximum precision.

S/N	Motif name	Consensus	Logo
1	Motif.5.13	CCTCC	
2	Motif.5.24	AGGAG	
3	Motif.6.6	CCTAGC	
4	Motif.9.12	TCGTTGGTT	
5	Motif.9.16	TTCCTCTCC	
6	Motif.10.12	ACATTGGATT	
7	Motif.10.14	AGCAACCTCG	
8	Motif.11.3	TATTTTTAGTA	
9	Motif.11.6	TTACCTTGTAT	

gene were regarded as one multi-dimensional vector, so that SVM algorithm can use them as input data. We used these data to train a SVM and to test the classification precision. We randomly selected motif combinations and tested their classification precision by SVM. Finally, the combination with a maximum precision of 0.89 containing 9 motifs was obtained (Table 2).

This motif combination can be regarded as the most discriminative ones in its class. In these motifs, we can still find some common features. For example, both motif 1 and motif 5 contained a sub consensus "CTT", while motif 4 and motif 6 contained consensus "TTC". We predicted that some of them should match part of the previous known motifs as well. To compare the function of these motifs, we matched them with entries in JASPAR database. The match was preceded using Stamp web-service, with some results listed in Table 3.

The core parameters to run the Stamp were: Metric = Pearson Correlation Coefficient, Alignment = Ungapped Smith-Waterman, Multiple Alignment method = Iterative Refinement, Tree = UPGMA, and Matching database: Jaspar v3 database.

Through this step, the discovered motifs were matched to function-known motifs. However, although most of them were successfully matched with the known motifs, some motifs were not. Possibly, they were new motifs, and further study is clearly needed to uncover their functions.

Function of selected motifs

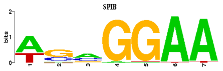
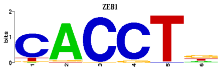
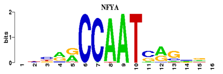
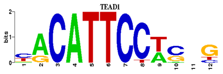
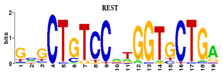
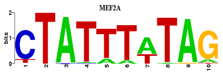
In order to identify the functions of these motifs, we queried all known motifs from the database and analyzed their functions. Binding motif of two factors match our discovered motif well, and the two factors were SPIB and ZEB1, respectively.

SPIB encodes hematopoietic-specific transcription factors and belongs to Winged Helix-Turn-Helix class. It is a B lymphocyte-specific ETS transcription factor. The ETS proteins share a conserved DNA binding domain and recognize various DNA target sites around a common core 5'-GGAA/T-3' (Ray-Gallet, Mao et al. 1995). SPIB binds DNA with a different affinity for certain sites and may activate unique target genes in B lymphocytes and interact with unique, though currently unidentified cofactors (Rao et al., 1999).

ZEB1 is a zinc-coordinating class Beta-Beta-Alpha zinc finger homeodomain enhancer-binding protein located in the 10p11.22. It encodes a TF that represses T-lymphocyte-specific inhibits interleukin-2 (IL-2) gene expression by binding to a negative regulatory domain near the IL2 transcription start site. It may be responsible for transcriptional repression of the IL-2 gene.

Both discovered factors are related to the immune system, but why their binding sites frequently occurred in our experiment is not known. One suspicious factor of KBD is selenium deficiency, and trace amounts of

Table 3. Motif matched from JASPAR database.

Motif Name	E value	Consensus alignment	Motif logo
SPIB	4.9083e-03	--CCTCC TTCCKST	
ZEB1	2.7587e-03	-AGGAG NAGGTG	
NFYA	1.8749e-03	----TCGTTGGTT--- NNNCYSATTGGYYNNN	
TEAD1	1.8247e-03	GGAGAGGAA---- -CNSWGGAATGTR	
REST	8.0433e-03	-------CGAGGTTGCT-- GNGCTGTCCNWGGTGCTGA	
MEF2A	2.5262e-10	TACTAAAAATA- --CTATAAATAG	

selenium are essential for the production of various cellular components such as enzymes. Selenium deficiency is a known trigger for several different autoimmune diseases, and impairs host innate immune response (Wang et al., 2009). Possibly, these genes are affected by the immune system change. In fact, the amount of selenium available in the soil varies from place to place which affects the amount of selenium found in local produces and food intake finally. Therefore, our finding may give the first molecular biological evidence that KBD is often an endemic disease.

Other factors also play a role in the disease through our research, such as MEF2A. MEF2A is a MADS box transcription enhancer factor 2 polypeptide A (myocyte enhancer factor 2A). In most cases, it binds to muscle-specific genes, but its functions in the molecular chemistry are unknown.

Conclusions

In this study, we used SVM selection method to identify motifs strongly associated with gene expression related to KBD. KBD is a special type of osteoarthritis with endemic distribution but no molecular biological evidence is reported. Through microarray experiments on blood samples of KBD patients and healthy individuals, we compared the transcript number of genes of the controls and cases, and 168 evidently differentially expressed genes were discovered.

The upstream sequences of these genes were used to discover motifs, and 186 motifs were discovered. After SVM selection, 9 motifs were finally screened out to represent the most relevant motifs. Two binding sites of two immune related factors were confirmed, showing that some genes were activated or suppressed by the changing immune system.

The possible mechanism of KBD, based on our study, can be deficiency of trace element intake, due to natural environmental degradation or food, which causes immune system to adapt, therefore affecting some gene expressions. Long term affected gene expressions in the cell may make KBD symptoms to occur consequently. Therefore, our finding may give the first molecular biological evidence that KBD is often an endemic disease. Furthermore, alternations of the living environment or the balance of trace element intake in food may help prevent this disease.

In comparison with the traditional method where lot of motifs are discovered, our method could reduce the total amount of motifs. So, we can focus on the most important ones to make further investigation. The methods used in this study could be applied to other microarray experiments to explore the underlying relationships between motif and gene functions.

ACKNOWLEDGEMENTS

This study was supported by the National Natural Science Foundation of China (No. 60601017, 30972556, 30630058, 41072020 and 05MY19) and the Scientific Research Foundation of Shaanxi Provincial Office of Health, P. R. China (No. 2010E06, 2010D21).

REFERENCES

Bailey TL, Elkan C (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2: 28-36.

Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares Jr. M, Haussler D (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. USA. 97(1): 262-267.

Chen X, Guo L, Fan Z, Jiang T (2008). W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. Bioinformatics, 24(9): 1121-1128.

Conlon EM, Liu XS, Lieb JD, Liu JS (2003). Integrating regulatory motif discovery and genome-wide expression analysis. Proc. Natl. Acad. Sci. USA. 100(6): 3339-3344.

FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C (2004) .Clustering of DNA sequences in human promoters. Genome Res. 14(8): 1562-1574.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5(10): R80.

He J, Hu HJ, Harrison R, Tai PC, Pan Y (2006). Rule generation for protein secondary structure prediction with support vector machines and decision tree. IEEE Trans. Nanobioscience, 5(1): 46-53.

Jaakkola T, Diekhans M, Haussler D (1999). Using the Fisher kernel method to detect remote protein homologies. Proc. Int. Conf. Intell. Syst. Mol. Biol., pp. 149-158.

Liu XS, Brutlag DL, Liu JS (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat. Biotechnol. 20(8): 835-839.

Lu CC, Yuan WH, Chen TM (2008). Extracting transcription factor binding sites from unaligned gene sequences with statistical models. BMC Bioinformatics, 9 Suppl 12: S7.

Mahony S, Benos PV (2007). STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. 35(Web Server issue): W253-258.

Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 38(Database issue): D105-110.

Rao S, Matsumura A, Yoon J, Simon MC (1999). SPI-B activates transcription via a unique proline, serine, and threonine domain and exhibits DNA binding affinity differences from PU.1. J. Biol. Chem. 274(16): 11115-11124.

Ray-Gallet D, Mao C, Tavitian A, Moreau-Gachelin F (1995). DNA binding specificities of Spi-1/PU.1 and Spi-B transcription factors and identification of a Spi-1/Spi-B binding site in the c-fes/c-fps promoter. Oncogene, 11(2): 303-313.

Seok J, Kaushal A, Davis RW, Xiao W (2010). Knowledge-based analysis of microarrays for the discovery of transcriptional regulation relationships. BMC Bioinformatics, 11 Suppl 1: S8.

Spyrou C, Stark R, Lynch AG, Tavare S (2009). BayesPeak: Bayesian analysis of ChIP-seq data. BMC Bioinformatics, 10: p. 299.

Wang C, Wang H, Luo J, Hu Y, Wei L, Duan M, He H (2009). Selenium deficiency impairs host innate immune response and induces susceptibility to *Listeria monocytogenes* infection. BMC Immunol. 10: p. 55.

Zamdborg L, Ma P (2009). Discovery of protein-DNA interactions by penalized multivariate regression. Nucleic Acids Res. 37(16): 5246-5254.