

A Data Mining Method to Predict Transcriptional Regulatory Sites Based on Differentially Expressed Genes in Human Genome

HSIEN-DA HUANG¹, HUEI-LIN CHANG⁴, TSUNG-SHAN TSOU³,
BAW-JHIUNE LIU⁴ AND JORNG-TZONG HORNG^{1,2,*}

¹*Department of Computer Science and Information Engineering*

²*Department of Life Science*

³*Institute of Statistics*

National Central University

Chungli, 320 Taiwan

⁴*Department of Computer Science and Engineering*

Yuan-Ze University

Chungli, 320 Taiwan

**E-mail: horng@db.csie.ncu.edu.tw*

Very large-scale gene expression analysis, i.e., UniGene and dbEST, is provided to find those genes with significantly differential expression in specific tissues. The differentially expressed genes in a specific tissue are potentially regulated concurrently by a combination of transcription factors. This study attempts to mine putative binding sites on how combinations of the known regulatory sites homologs and over-represented repetitive elements are distributed in the promoter regions of considered groups of differentially expressed genes. We propose a data mining approach to statistically discover the significantly tissue-specific combinations of known site homologs and over-represented repetitive sequences, which are distributed in the promoter regions of differentially gene groups. The association rules mined would facilitate to predict putative regulatory elements and identify genes potentially co-regulated by the putative regulatory elements.

Keywords: regulatory site, transcription factor, data mining, gene expression, UniGene, EST

1. INTRODUCTION

Very large-scale gene expression analysis, i.e., UniGene [1] and dbEST [2], is provided to find those genes with significantly differential expression in specific tissues. Single-pass partially sequencing of cDNA clones from diverse tissues are deposited in the form of expressed sequence tags (ESTs) in dbEST [3-5]. The sequences in GenBank including ESTs are partitioned into a non-redundant set of gene-oriented clusters and are stored in the UniGene database. Each of UniGene clusters means as a unique gene with numerous sequences from different EST libraries, which contain information such as the tissue type, organism protein similarity modeling, and the LocusLink identifier. It has also been used for gene mapping and large-scale expression analysis.

Bortoluzzi et al. proposed a computational approach to large-scale analysis of gene

expression in several human adult tissues [6]. The basic assumption of their method [6] is that the level of activity and the tissue expression pattern of a given gene may be inferred from the number of corresponding ESTs obtained from unbiased cDNA libraries from the considered tissues. A computational and statistical study on a large set of gene expression data pertaining six adult human tissues, uterus, ovary, brain, liver, skeletal muscle, retina, was performed for analyzing the expression of ribosomal protein genes. While Bortoluzzi et al. considered only the ribosomal protein genes for which expression levels were in statistically different according to both R statistics [7] and the AC test [8], we attempt to select all genes that are differentially expressed in the same considered tissues for further analysis and hope to discover the associations of known site homologs and oligonucleotides to predict potentially regulatory sites.

Many experimental identifying TF binding sites have been collected in TRANSFAC [9], which is the most complete and well maintained database on transcription factors, their genomic binding sites and DNA-binding profiles [9]. Our approach takes the known TF site homologs in the considered gene promoter region into account in order to correlate the over-represented repetitive oligonucleotides located in upstreams of considered differentially expressed gene group.

Brázma et al. [10] developed a general software tool to find and analyze combinations of TF binding sites that occur often in gene upstream regions in the yeast genome. Horng et al. [11-13] also proposed a data mining approach, association rules, to investigate combinations of known site homologs and over-represented repeats in yeast genome, and some significant oligonucleotides are predicted as putative sites based on their significant correlation to known sites homologs.

Moreover, we are interested in the potentially co-regulated genes, which are the differentially expressed genes in specific tissues discovered by computational and statistical approaches. This study attempts to find differentially expressed genes by statistically analyzing EST libraries from six adult human tissues, i.e. uterus, ovary, brain, liver, skeletal muscle, and retina. The differentially expressed genes in a specific tissue are potentially regulated concurrently by combinations of transcription factors. Based on this idea we apply the data mining approach proposed in [11, 13] to mine putative binding sites on how combinations of the known regulatory site homologs and over-represented repetitive oligonucleotides are distributed in the promoter regions of considered differentially expressed gene groups. Some interesting associations of known site homologs and over-represented repetitive oligonucleotides are found in ovary, skeletal muscle, and liver.

2. METHODS

The approach initially observed genes with highly and differentially expression in the same tissues, while they are not in other tissues [6] by calculating the R statistic and P_{AC} values from gene expression data partitioned into different tissue categories. Moreover, we computationally identify the combinations of known transcription factor (TF) site of *Homo sapiens* from TRANSFAC [9] and repetitive sequences from RSDB [12] located in the promoter regions of groups of genes which are highly and differentially expressed in a specific tissue. Then, a data mining approach is applied to mine the asso-

ciation rules in the combinations of the known site homologs and over-represented repetitive oligonucleotides. A Chi-square test is then applied to select certain significant rules. Finally, the over-represented repetitive oligonucleotides in the association rules are candidates of putative regulatory elements [11-13].

2.1 Materials

The sequences of promoter regions, i.e., upstream regions from – 1 bps to – 2000 bps (where – 1 bps is the position of the translational start site), are extracted directly from the draft sequences of the target genome, i.e., *Homo sapiens*, and the gene annotations are obtained from GenBank [14]. The differential gene expression analysis involved six human tissues, uterus, ovary, brain, liver, skeletal muscle and retina, for which a sufficient number of ESTs, obtained from unbiased cDNA libraries, was available in the UniGene database (build #147). The number of cDNA libraries, UniGene clusters and EST sequences retrieved for each tissue are shown in Table 1. For example, unbiased cDNA libraries consisting of library IDs 119, 312, 732, 733, and 3600 are selected for the uterus, as well as library IDs 24, 272, and 500 which are selected for the skeletal muscle. In uterus, 39340 EST sequences and 5957 annotated genes from 11628 UniGene clusters are analyzed, as well as 2790, 1855, 844, 2236, and 2509 genes in ovary, brain, liver, skeletal muscle, and retina, respectively.

Table 1. The six human tissues in this study.

Tissue	UniGene cDNA Libraries (Lib. #)	Number Retrieved		
		ESTs	UniGene Clusters	Human Genes (Annotated)
Uterus	119, 312, 732, 733, 3600	39,340	11,628	5,957
Ovary	576, 652, 935, 1369, 1380, 3223, 3225	14,312	4,384	2,790
Brain	128, 859, 857, 15, 255	4,953	2,945	1,855
Liver	155	6,270	1,329	884
Skeletal muscle	24, 272, 500	21,366	3,795	2,236
Retina	177, 178, 228, 313	14,001	4,812	2,509

As shown in Fig. 1, 17976 human annotated genes with gene symbol are stored in GenBank (Sept. 6, 2001), as well as 96,105 clusters in UniGene database. 15060 genes of GenBank are selected because they occur exactly once in the draft sequences of the human genome. Also, 15198 annotated UniGene clusters are selected. By cross-referencing the gene symbol of these genes selected from GenBank and UniGene, 12060 are found with the same gene symbol.

The experimental identifying transcription factor binding sites can be obtained from TRANSFAC [9]. TRANSFAC database [9] (professional 5.4) contains a total of 11,537 site sequences, while the number of binding sites of *Homo sapiens* is 1269. Most of the known sites used in this study are consensus patterns. The data in TRANSFAC has the following features. A transcription factor binding site accession number may have

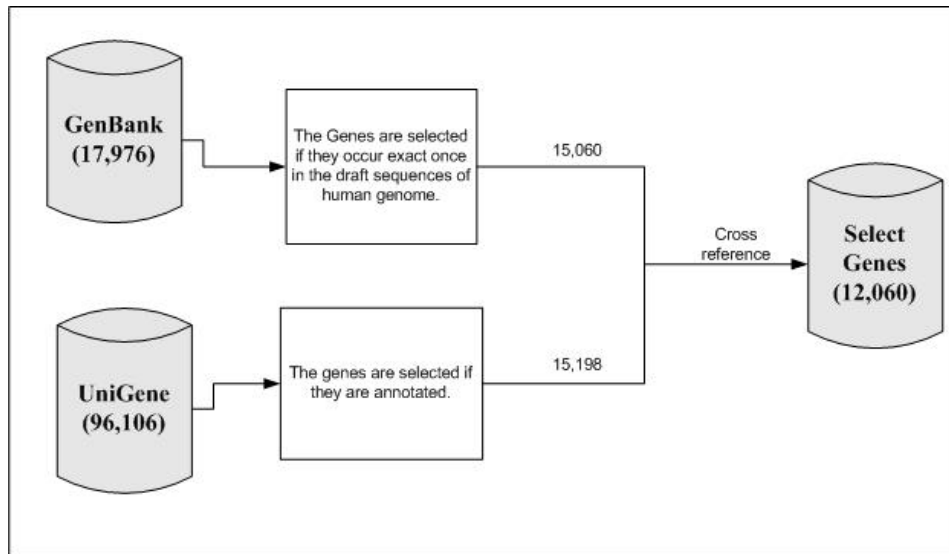


Fig. 1. The selecting process of genes from GenBank and UniGene.

different consensus sequences. Different binding site accession numbers may have a same consensus sequence. Wild characters such as 'M' or 'W' used in TRANSFAC cause the sequences to cover other sequences. Small consensus sequences may appear in larger ones.

The repetitive sequences of the human genome are generated from the merging draft sequences of the human genome from GenBank and stored in the repetitive sequence database (RSDB) [12], where the whole human genome sequences are merging of nucleotide contigs without assembling them for the reason of only considering the site occurrence in the whole human genome.

2.2 Differential Gene Expression

We select six adult human tissues, brain, liver, skeletal muscle, ovary, retina and uterus, which have also been analyzed in [6], to perform a computational and statistical analysis. The aim of this step is to find the genes with differential expression by computing their R value [7] and P_{AC} value [8] based on the amounts of EST sequences corresponding in each considered tissue. Each of the 12,060 UniGene clusters are analyzed to be differentially expressed in any one of the six tissues when the calculated R value is greater than 35.61 (the value is associated with a probability value smaller than 0.0001) and P_{AC} value is smaller than 0.0001. Genes showing (are expressed) in a given tissue at a level of expression significantly higher than in all the other tissues were considered as 'differentially expressed'. All genes which are significantly more expressed in a specific tissue than in others are identified and then grouped by tissues.

The R-statistic is denoted as R_j for gene j in Eq. (1), where m is the number of cDNA libraries, $x_{i,j}$ is the number of transcript copies of gene j in the i th library, and N_i is the total number of cDNA clones sequenced in the i th library. The frequency f_j of gene transcript copies of gene j in the i th library over all libraries is given by Eq. (2).

$$R_j = \sum_{i=1}^m x_{i,j} \ln \left(\frac{x_{i,j}}{N_i f_j} \right) \tag{1}$$

$$f_j = \frac{\sum_{i=1}^m x_{i,j}}{\sum_{i=1}^m N_i} \tag{2}$$

For instance, the number of genes in the six adult human tissues, uterus, ovary, brain, liver, skeletal, and retina, are 39340, 14312, 4953, 6270, 21366 and 14001, respectively. Furthermore, the numbers of transcript copies of gene ALB in six adult human tissues are 0, 0, 0, 896, 0, and 0, respectively. By applying Eqs. (1) and (2), the R-value of gene ALB is 2487.296 as shown in the following calculations. The larger the R-value computed from different tissues, the more differential tissues there are among the select tissues. Therefore, the ALB gene is expressed more differential (or preferable) in liver than the other tissues.

$$f_{ALB} = \frac{0 + 0 + 0 + 896 + 0 + 0}{39,340 + 14,312 + 4,953 + 6,270 + 21,366 + 14,001} \approx 0.0089$$

$$R_{ALB} = 0 + 0 + 0 + (896 \times \ln \frac{896}{6,270 \times 0.0089}) = 2487.296$$

Another gene expressed difference measure, P_{AC} , is calculated to investigate the differential gene expression among two tissues. To establish the probability for a given cDNA library, e.g., “Liver” to be picked up x times when the sampling size was N_1 and another cDNA library, e.g., “Brain” to be picked up y times when the sampling size was N_2 , the expected probability of observing y occurrences of a clone already observed x times is given by the simple Eq. (3) which is established in [8].

$$p(y | x) = \left(\frac{N_2}{N_1} \right)^y \frac{(x + y)!}{x! y! \left(1 + \frac{N_2}{N_1} \right)^{(x+y+1)}} \tag{3}$$

For instance, the total number of ESTs in uterus and skeletal muscle are 39,340 and 21,366, respectively. The number of transcript copies of RPS25 in uterus and skeletal muscle are 25 and 154, respectively. The P-value is $4.62 * E^{-52}$ as shown in the following example by applying Eq. (3). The P-value is very small and means that the RPS25 gene is more differential. To extend the calculation to handle multiple libraries, Bonferroni correlation is applied in this study.

$$p(uterus | skeletal - muscle) = \left(\frac{39,340}{21,366} \right)^{25} \frac{(154 + 25)!}{154! 25! \left(1 + \frac{39,340}{21,366} \right)^{(154+25+1)}} = 2.31 \times 10^{-45}$$

2.3 Preprocessing and Mapping

The TF binding sites categorized in *Homo sapiens* from TRANSFAC and repetitive oligonucleotides in RSDb are first prepared. For each group of considered differentially expressed genes in a specific tissue, all of the known human regulatory sites are directly located into the gene promoter regions from -1 to -2000 bps, as well as the repetitive oligonucleotides are located. The occurrences of each known site homologs and over-represented repetitive oligonucleotides are calculated and then provided to the following statistical analysis.

2.4 Detecting Over-represented Repetitive Oligonucleotides

To detect the over-represented oligonucleotides in upstream regions, oligo-analysis has been described before and is based on a systematic counting of occurrences for all the possible oligonucleotides of a given sequences [15]. An advantage of the method is that it is able to detect all the over-represented patterns of a given length in a single run. Here, we perform a statistical method to discover statistical significant oligonucleotides, i.e., small length of DNA sequences, within the upstream regions of genes by comparing their occurrence frequencies to the background occurrence frequencies in whole human genome, where the occurrence frequencies of oligonucleotides are obtained from *i*-Human.

Based on the concept addressed above, we attempt to test the hypothesis of whether an oligonucleotide is over-represented in gene upstream regions.

Nucleotide succession is not random, and some oligonucleotides are clearly over-represented, notably the poly (A), poly (T), and poly (AT) chains. An additional bias results from the fact that oligonucleotides are differently represented in coding regions versus non-coding sequences [15]. A specific expected frequency has to be used this way for each oligonucleotide sequence. van Helden *et al.* proposed a statistical method to estimate the probability of observing exactly n occurrences of the oligonucleotide b within the promoter regions of a gene family by the binomial equation. The values with the highest probability are the most over-represented oligomers. The advantage of the significance value is that its threshold can be selected and its values interpreted independently of oligonucleotide size, upstream sequence size, and number of genes within the family. The over-represented repetitive oligonucleotides in gene upstreams are obtained by applying the statistical method in [15]. The repetitive oligonucleotides, which are with significant values exceeding the threshold, are selected as significant over-represented ones.

Then a specific expected frequency has used for each repetitive oligonucleotide to determine the statistical significance.

$$T = 2 \times S \times (L_i - w + 1) \quad (4)$$

$$P(\text{occ}\{b\} = n) = \frac{T!}{(T-n)!n!} \times (F_e\{b\})^n \times (1 - F_e\{b\})^{(T-n)} \quad (5)$$

$$P(occ\{b\} \geq n) = \sum_{j=n}^T P(occ\{b\} = j) = 1 - \sum_{j=0}^{n-1} P(occ\{b\} = j) \quad (6)$$

where $F_c\{b\}$ is the frequency observed throughout all non-coding segments of the whole yeast genome; T represents the total number of possible matching positions for a pattern of length w across both strands of the sequence set; S is the number of sequences in the set; L_i is the length of the i th sequence in the set; $P(occ\{b\} = n)$ is the probability to observe exactly n occurrences of the oligomer b ; $P(occ\{b\} \geq n)$ is the probability to observe n or more occurrences of the oligomer b .

$$D = 4^w - (4^w - N_{pal}) / 2 \quad (7)$$

D is the distinct number of oligomers; N_{pal} is the palindromic oligomers. Last we defined a significance coefficient:

$$sig = -\log_{10}[P(occ\{b\} \geq n) \times D] \quad (8)$$

which the highest values for this parameter correspond to the most over-represented sequences.

Another significance measure is associated with the frequency of an oligonucleotide occurrence within in upstream regions as compared to all human non-coding sequences, as a background. Here, we detect the over-represented oligonucleotides occurring in the upstream regions of selected genes. If $F_c\{b\}$ is the occurrence probability of oligonucleotide b observed in all non-coding regions of the human genomic sequence, we would expect the b oligonucleotide also occurs $u = T \times F_c\{b\}$ times in the upstream regions of genes, where T represents the total number of possible matching positions for a oligonucleotide of length w across both strands of the sequence set. Using a simple binomial model, the standard deviation becomes $\sigma = \sqrt{T \times F_c\{b\} \times (1 - F_c\{b\})}$. Let n is the amount of the considered oligonucleotide b occurring in upstream regions, the Z-score is calculated as $Z = (n - u) / \sigma$. The probability of observing at least n successes, as given by Chebyshev's Theorem, is less than or equal to $p_value = ((n - u) / \sigma)^{-2}$. If $Z > 0$, the lower the p-value, this repeat is the over-represented repetitive oligonucleotides, on the other hand, if $Z < 0$, the lower the p-value, the repeat is the under-represented oligonucleotides.

We present an example to explain how to find over-represented repetitive sequences. The background probability of a site sequence "ATCTAG" occurring in whole genome, i.e., the size is about 2.8 billions bps by merging the nucleotide contigs as mentioned, is $2.51 * 10^{-4}$. To consider there ten genes with higher expression in liver and the number of occurrence in those genes is fifty times. The p-value and Z-score are computed as following example where $F_c\{b\}$ is the background probability of the site and T is the total upstream lengths of considered genes. Then, the standard deviation σ is about 3.16 and Z value is 12.65 (the p_value is 0.0062). The repetitive sequences are selected as significant over-represented repetitive oligomers if their *sig* coefficients are larger than 0 and Z-score are greater than 4.47 (p-value = 0.05).

2.5 Mining Sites Associations and Pruning

In the following we describe how to mine associations from the combinations of the

transcription factor binding sites and over-represented repetitive sequences. Consider a large database of transactions, where each transaction consists of a set of items. An association rule is an expression of the form $A \Rightarrow B$, where A and B are the sets of items. The mining of an association rule is that a transaction in the database that contains A also tends to contain B . For example, 90% of the people who purchase beer also purchase diapers. Herein, 90% is called the confidence of the rule. The support of the rule $A \Rightarrow B$ given herein is the percentage of transactions that contain both A and B .

The formal statement of the problem is described as follows. Let $S = \{s_1, s_2, \dots, s_m\}$ be a set of known sites homologs from TRANSFAC and $R = \{r_1, r_2, \dots, r_n\}$ be a set of over-represented repetitive sequences in human from RSDB [12]. The union of the sets S and R is called 'item set'. Let $G = \{g_1, g_2, \dots, g_m\}$ be a group of genes with differential expression in a specific tissue. Each promoter region of a gene is mapped to a transaction containing a set of known regulatory site homologs and over-represented repetitive oligonucleotides, also called items.

Assume that a promoter region S contains A , a set of items of I , if $A \subseteq S$. An association rule is an implicate of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \cap B = \emptyset$. The rule $A \Rightarrow B$ holds in the set of promoter regions D with confidence $c\%$ if $c\%$ of transactions in D contains A and also B . The rule $A \Rightarrow B$ has support value $s\%$ in the gene upstream regions if $s\%$ of promoter regions in D contained $A \cup B$. In our experiments, the minimum support is set to 40%. The association rules are generated if the rule has a higher support and confidence than the threshold specified by users. The Apriori algorithm [16] is implemented and applied here to mine association rules.

Fig. 2 presents an example of a mapping between the gene promoter regions and regulatory sites, i.e., known TF binding sites and over-represented repetitive oligonucleotides. The GID denotes the gene upstreams, and RID denotes the regulatory sites. For example, the APOH contains the regulatory sites $RID\{1, 3, 4\}$. L_i denotes the phase of discovering combinations of length i . The combination of sites $\{2, 3, 5\}$ in L_3 with support 2, i.e., two genes of PLG and HP contain the combinations.

GID		RID	C ₁		L ₁	
GID	RID	GID	RID	Item set	Sup	
APOH	1, 3, 4	APOH	{ {1}, {3}, {4} }	{ 1 }	2	
PLG	2, 3, 5	PLG	{ {2}, {3}, {5} }	{ 2 }	3	
HP	1, 2, 3, 5	HP	{ {1}, {2}, {3}, {5} }	{ 3 }	3	
AMBP	2, 5	AMBP	{ {2}, {5} }	{ 5 }	3	

C ₂		C ₂		L ₂	
Item set	GID	RID	Item set	Sup	
{ 1 2 }	APOH	{ {1 3} }	{ 1 3 }	2	
{ 1 3 }	PLG	{ {2 3}, {2 5}, {3 5} }	{ 2 3 }	2	
{ 1 5 }	HP	{ {1 2}, {1 3}, {1 5}, {2 3}, {2 5}, {3 5} }	{ 2 5 }	3	
{ 2 3 }	AMBP	{ {2 5} }	{ 3 5 }	2	
{ 2 5 }					
{ 3 5 }					

C ₃		C ₃		L ₃	
Item set	GID	RID	Item set	Sup	
{ 2 3 5 }	PLG	{ {2 3 5} }	{ 2 3 5 }	2	
	HP	{ {2 3 5} }			

Fig. 2. An illustrative example of the association of transcriptional regulatory sites.

A huge number of combinations are found in each the considered tissues with enough differential expressed genes. The Chi-square test is then used to investigate the site correlations in each combination. However, the sites in any combinations depend significantly on the differential expression genes in a specific tissue when the calculation of the Chi-square value exceeds a threshold of 3.84 (with degree of freedom 1 and $\alpha = 0.05$).

The Chi-square test (χ^2) is a widely used method for testing the independence and (or) correlation, and is applied to discover the significant associations rules in [17]. Let f_o be an observed frequency, and f be an expected frequency. The Chi-square test is used to measure the significance of the deviation from the expected values. The value of χ^2 is defined as $\chi^2 = \sum ((f_o - f)^2 / f)$. As the example in Fig. 3 shows, we would like to test the correlation of the sites in the combination “aaatat, ttgaa”. The two oligonucleotides, “aaatat” and ”ttgaa”, occur concurrently in 23 gene promoter regions of 35 genes and totally in a specific tissue, e.g., Skeletal muscle, while six genes do not contain any of the two sequences. Five genes contain the sequence “ttgaa”, but do not contain “aaatat”. The four conditions are constructed as a 2 by 2 contingency table shown in Fig. 3.

	ttgaa	$\overline{\text{ttgaa}}$	Row Total
aaatat	23	1	24
$\overline{\text{aaatat}}$	5	6	11
Column Total:	28	7	35

Fig. 3. A 2-by-2 contingency table to show the number of genes containing sites “aaatat” and “ttgaa” in promoter regions.

If the correlation of the two sites indicates they are independent, we expect the total number of the site “aaatat”, e.g., 24, is divided into the ratio of 80% (28/35) and 20% (7/35). If it is higher than a certain threshold, for example, 3.84 is at the 95% significance level with 1 degree of freedom, then we reject the hypothesis of independence of site occurrences. We say that the site occurrences in the combination are correlated. As the example given in Fig. 3, the chi-square value is 11.965, which exceeds 3.84, the occurrence of the two sites are dependent.

2.6 Tissue-specific Combinations

After the discovery of significant associations of known site homologs and over-represented repetitive oligonucleotides, we propose here a D-value statistic to determine the association which is preferable to some specific tissues, i.e., the association is more frequently found in the specific tissues. In order to investigate the occurrence differences of the site association mined from different tissues, we propose a D-value statistic to compute the hypothesis that each frequency of occurrence is consistent with the others. Computing the D-value can extract the combinations whose the occurrence varies most across different tissues. The statistic is denoted as D_j for combination j in Eq. (9), where m is the number of tissues, $x_{i,j}$ is the number of genes containing the combination j

in tissue i , and N_i is the total number of considered genes in tissue i . The frequency f_j of the combination j in tissue i over all tissues is given by Eq. (10).

$$D_j = \sum_{i=1}^m X_{i,j} \ln\left(\frac{X_{i,j}}{N_i f_j}\right) \quad (9)$$

$$f_j = \frac{\sum_{i=1}^m x_{i,j}}{\sum_{i=1}^m N_i} \quad (10)$$

For instance, the number of the considered genes (differentially expressed) in the tissues of ovary, liver and skeletal muscle are 10, 21 and 35, respectively. The site combination, “tataca, ttgaaa”, can be located in 2, 3 and 19 gene promoter regions of ovary, liver, and skeletal muscle, respectively. By applying Eq. (7), the D-value in this example is calculated to be 8.23. The greater D-value is in different tissues, the more tissue-specific the combination among the select tissues is.

3. RESULTS

We first analyze the UniGene clusters and cDNA libraries as mentioned in section on materials. The number of EST sequences corresponding to each differential expression gene in each considered tissue was obtained and merged in a matrix of 47 rows (genes) and 6 columns (tissues) (Table 2, columns 3-8). For example, Hs.77039 is a higher expressed in uterus in the other tissues; ten genes are a higher expression in ovary; three genes are higher expression in brain; 21 genes are higher expression in liver; 35 genes are higher expressed in skeletal muscle; and four genes are higher expressed in retina. In this study we adopt ten genes in ovary, twenty-one genes in liver and 35 genes in skeletal muscle for further analyses and the potentially co-regulated genes are used for the prediction of regulatory elements. The UniGene ID is given in the first column, and the gene symbol is given in the second one. The following six columns show the amount of EST sequences from different cDNA libraries, i.e., from different tissues, that can be assigned into the cluster of each row. The amounts of EST sequences corresponding to each UniGene clusters are shown in columns 3-8 (tissues). The value of R statistic of each UniGene cluster is given in the R column, as well as the P_{AC} value is given in the last column. The differentially expressed genes, are marked with one, two or three asterisks (P_{AC} value below 0.01, 0.001 or 0.0001, respectively). For instance, the UniGene cluster Hs.929 with the gene symbol “MYH7” is found to be highly expressed in skeletal muscle because of the high R value [7] and low P_{AC} value [8].

The genome sequences and gene information are obtained from NCBI. We use the R statistic and P_{AC} value to analyze genes expressions in six adult human tissues. Since less than 6 (specified threshold) the numbers of differential expressed genes are found in the tissues of uterus, brain, and retina, these three tissues are not analyzed further for prediction of regulatory sites. We just focus on ovary, liver and skeletal muscle. Table 3

Table 2. The UniGene clusters which are differentially expressed.

UniGene ID	Genes	Uterus	Ovary	Brain	Liver	Skeletal muscle	Retina	R	Expressed Tissue Favors
Hs.71	AZGP1	0	0	0	22	0	0	60.98	Liver*
Hs.929	MYH7	0	0	0	0	318	0	491.56	Skeletal muscle***
Hs.931	MYH2	0	0	0	0	200	0	309.16	Skeletal muscle***
Hs.1252	APOH	0	0	0	71	0	0	196.80	Liver***
Hs.1288	ACTA1	0	0	0	0	268	0	414.27	Skeletal muscle***
Hs.1940	CRYAB	3	0	0	0	90	7	117.09	Skeletal muscle***
Hs.2186	EEF1G	131	94	3	10	46	27	38.57	Ovary**
Hs.2257	VTN	0	0	0	28	0	0	77.61	Liver**
Hs.3462	COX7C	7	2	0	2	69	2	75.16	Skeletal muscle*
Hs.29797	RPL10	57	17	0	4	83	5	46.02	Skeletal muscle*
Hs.31439	SPINT2	43	69	0	0	0	1	96.19	Ovary***
Hs.69771	BF	16	1	0	38	0	0	84.43	Liver***
Hs.73454	TNNT3	0	0	0	0	101	0	156.12	Skeletal muscle***
Hs.73980	TNNT1	2	3	0	0	354	0	525.22	Skeletal muscle***
Hs.74335	HSPCB	92	67	0	4	28	7	47.73	Ovary*
Hs.74565	APLP1	7	0	15	0	0	0	37.90	Brain*
Hs.75535	MYL2	0	0	0	0	91	0	140.67	Skeletal muscle***
Hs.75576	PLG	0	0	0	34	0	0	94.24	Liver***
Hs.75990	HP	2	2	0	229	0	0	617.51	Liver***
Hs.76067	HSPB1	20	15	1	0	77	0	70.75	Skeletal muscle*
Hs.76177	AMBP	0	0	0	41	0	0	113.64	Liver***
Hs.76452	CRP	0	0	1	45	0	0	122.92	Liver***
Hs.76669	NNMT	6	1	0	28	0	0	64.78	Liver**
Hs.77039	RPS3A	204	160	1	1	22	1	189.10	Uterus***
Hs.77899	TPM1	9	3	0	2	97	1	111.44	Skeletal muscle***
Hs.79933	CCNI	16	5	1	0	8	82	103.41	Retina***
Hs.80617	RPS16	136	90	1	0	45	0	94.58	Ovary**
Hs.83760	TNNI2	0	0	0	0	131	0	202.50	Skeletal muscle***
Hs.83870	NEB	0	0	0	0	79	0	122.12	Skeletal muscle**
Hs.84673	TNNI1	0	0	0	0	240	0	370.99	Skeletal muscle***
Hs.93194	APOA1	0	0	0	275	0	0	762.25	Liver***
Hs.99858	RPL7A	93	80	0	2	22	1	78.63	Ovary***
Hs.108124	RPS4X	65	13	0	0	171	6	143.87	Skeletal muscle***
Hs.112318	LOC54543	5	2	1	0	81	1	99.16	Skeletal muscle*
Hs.113029	RPS25	25	5	0	6	154	3	156.26	Skeletal muscle***
Hs.114346	COX7A1	1	0	0	0	77	0	114.61	Skeletal muscle**
Hs.118804	ENO3	0	0	0	0	145	1	220.13	Skeletal muscle***
Hs.118836	MB	1	0	0	0	388	0	593.74	Skeletal muscle***
Hs.118845	TNNC1	1	0	0	0	84	1	121.86	Skeletal muscle***
Hs.151242	SERPING1	7	3	1	60	4	19	122.28	Liver***
Hs.158295	MYL1	0	0	0	0	94	0	145.30	Skeletal muscle***

Hs.159154	TUBB4	3	4	19	0	0	0	47.81	Brain**
Hs.169849	MYBPC1	0	0	0	0	120	0	185.49	Skeletal muscle***
Hs.172004	TTN	6	0	0	0	237	0	343.83	Skeletal muscle***
Hs.177486	APP	9	5	1	0	0	53	76.81	Retina*
Hs.177592	RPLP1	16	2	1	9	113	2	116.84	Skeletal muscle*
Hs.182421	TNNC2	0	0	0	0	221	0	341.62	Skeletal muscle***
Hs.182426	RPS2	139	152	1	4	42	0	148.76	Ovary***
Hs.183706	ADD1	9	3	25	0	4	15	49.45	Brain*
Hs.184411	ALB	0	0	0	896	0	0	2483.54	Liver***
Hs.194366	TTR	0	0	0	64	0	14	168.25	Liver***
Hs.198246	GC	0	0	0	53	0	0	146.91	Liver***
Hs.198281	PKM2	70	149	4	0	46	18	126.47	Ovary***
Hs.231581	MYH1	0	0	0	0	73	0	112.84	Skeletal muscle**
Hs.234234	ALDOB	0	0	0	46	0	0	127.50	Liver***
Hs.234726	SERPINA3	18	1	0	79	0	19	171.14	Liver***
Hs.237658	APOA2	1	0	0	45	0	0	120.85	Liver***
Hs.238756	MYOZ1	0	0	0	0	103	0	159.22	Skeletal muscle***
Hs.247565	RHO	0	0	0	0	0	76	149.60	Retina***
Hs.252259	RPS3	96	107	0	7	21	1	103.98	Ovary***
Hs.254105	ENO1	138	263	5	0	4	9	328.88	Ovary***
Hs.279604	DES	1	0	0	0	101	1	147.78	Skeletal muscle***
Hs.279860	TPT1	61	10	0	4	94	6	58.04	Skeletal muscle**
Hs.284176	TF	0	0	0	79	1	36	214.16	Liver***
Hs.284394	C3	30	5	0	68	1	2	135.83	Liver***
Hs.296290	RPL37A	24	7	0	0	325	1	411.71	Skeletal muscle***
Hs.297681	SERPINA1	18	5	0	149	1	0	355.26	Liver***
Hs.300772	TPM2	7	0	0	0	328	0	479.56	Skeletal muscle***
Hs.334347	CKM	0	0	0	0	573	1	880.35	Skeletal muscle***
Hs.334842	K-ALPHA-1	57	65	0	1	2	9	70.06	Ovary***
Hs.336920	GPX3	1	5	3	4	4	76	116.60	Retina**
Hs.337445	RPL37	4	8	2	2	129	0	160.52	Skeletal muscle***
Hs.343603	TCAP	2	0	0	0	156	0	232.29	Skeletal muscle***
Hs.346935	HPX	0	0	0	70	0	0	194.03	Liver***

Table 3. The associations of known site homologs and OR repeats mined in each considered tissue.

Tissue Type	Amount						
	Genes	Over-represented Repeats	Known Site homologs	Average Sites	Maximum Sites	Site Associations (before pruning)	Significant Site Associations
Ovary	10	61	158	59.10	69	5,573	3,941
Liver	21	87	186	76.76	94	1,711	1,042
Skeletal muscle	35	70	220	68.41	89	13,250	8,766

shows the numbers of known site homologs and the over-represented repetitive elements in promoter regions that are related to genes with differential expression in the specific tissue categories, i.e., ovary, skeletal muscle or liver. The minimum support is set to 60% and confidence is set to 80%, e.g., each of the found associations is contained in six of ten considered genes. “Average” and “Maximum” column indicate the average and maximum amounts of known sites or over-represented (OR) repeats in the promoter regions, respectively. For example, 5,573 associations are discovered in 10 promoter regions in ovary, where on 158 known site homologs and 61 over-represented repeats are located in the promoter region. The maximum number of known site homologs or over-represented oligonucleotides in ovary is 69. Finally, 3941 significant associations are left after applying Chi-square tests.

Table 4. Significant site associations.

Tissue Type	Site Combinations	Conf	Sup	χ^2	Site Combinations (with identifier of known site)
Ovary (10 genes)	acaggc, CATTT, GGTTA => agctga	1.00	0.70	6.85	acaggc, H\$GMCSF_03, H\$WT1_04 => agctga
	caagca, GCCCC => TTCCTT	0.87	0.70	5.83	caagca, H\$DPOLB_04 => H\$CDC2_02
	accagc, CATCTG => CTGTC	0.87	0.70	5.83	accagc, H\$IGKL_11 => H\$GG_12
	agacag, CTGTC => CATCTG	1.00	0.70	10.0	agacag, H\$GG_12 => H\$IGKL_11
	GCCCC, GGTGGG => ac-cagc	1.00	0.70	10.0	H\$DPOLB_04, H\$GPB_02 => accagc
Liver (21 genes)	TATAA => ctcagc	0.94	0.67	5.21	H\$ASCC_04 => ctcagc
	aggaga, CACCC, GTCAC => ggagaa	1.00	0.71	12.3	aggaga, H\$GG_13, H\$CLASE_04 => gcctcc
	aacaag, CTAAT => TAAAT	1.00	0.62	3.86	aacaag, H\$GG_22 => H\$GRH_03
	AAGTGA => agcaca	0.86	0.62	9.45	H\$IFNB_02 => agcaca
	CTGTC, GTCAC => aactga	0.93	0.78	13.0	H\$GG_12, H\$CLASE_04 => ggctga
Skeletal muscle (35 genes)	CACCC, TGGCA => agctgg	0.88	0.65	4.33	H\$GG_13, H\$ALBU_03 => agctgg
	GTCAC, GCCCC => gcctcc	1.00	0.60	7.35	H\$CLASE_04, H\$DPOLB_04 => gcctcc
	CATCTG => aggcag	0.91	0.60	5.35	H\$IGKL_11 => aggcag
	caggag, GGTGGG => agcctg	0.95	0.60	8.84	caggag, H\$GPB_02 => agcctg
	ctggtc, GGGCA => GCCCC	0.95	0.60	8.84	ctggtc, H\$CATHD_01 => H\$DPOLB_04

Several interesting and significant associations in tissues “Ovary”, “Liver” and “Skeletal muscle” are shown in Table 4. The first column in Table 4 is the tissue type and the second column are the associations containing known site homologs in uppercase letters and over-represented repetitive oligomers in lowercase; the third column is the confidence of the association; the fourth column is the support value; the fifth column is the Chi-square value; the last column is similar to the second one except for the use of transcription factor names in TRANSFAC [9]. For instance, the association of “acaggc,

CATTT, GGTTA => agctga” is discovered in ovary, where the support value is 0.70, the confidence value is 1.0, and the χ^2 value is 6.85. “CATTT” and “GGTTA” are known sites with TRANSFAC symbol of “YY1” and “WT1”, respectively, and “acagc” and “agctga” are a significant over-represented repetitive oligonucleotide. Similarly, “TAAAT => ctacg” is a site association found in liver with support value, confidence value, and χ^2 value of 0.67, 0.94, and 5.21, respectively; “CACCC, TGGCA => agctgg” is a site association found in skeletal muscle with support value, confidence value, and χ^2 value of 0.65, 0.88, and 4.33, respectively.

Table 5. An example of occurrences of the association.

Genes	UniGene ID (Hs.#)	Occurrences of Sites Association “aacaag, HS\$GG_22 => HS\$GRH_03”
APOH	Hs.1252	[-1800]-TAAAT-[94]-CTAAT-[224]-%ATTAG-[67]-aacaag-[51]-%ATTAG-[37]-%ATTAG-[7]-%ATTTA-[107]-%ATTTA-
PLG	Hs.75576	[-1999]-CTAAT-[9]-TAAAT-[101]-%ATTAG-[285]-%ATTAG-[8]-CTAAT-[21]-%ATTTA-[171]-TAAAT-[20]-%ATTAG-[266]-%cttggt-
HP	Hs.75990	[-1778]-CTAAT-[193]-%ATTTA-[146]-%cttggt-[127]-CTAAT-[189]-%ATTTA-[1]-CTAAT-
AMBP	Hs.76177	[-1999]-%cttggt-[227]-%ATTTA-[94]-%ATTTA-[6]-TAAAT-[135]-%ATTAG-[153]-TAAAT-[3]-TAAAT-[26]-%ATTAG-[287]-%ATTTA-
NNMT	Hs.76669	[-1933]-CTAAT-[37]-%ATTTA-[85]-%ATTTA-[56]-TAAAT-[45]-TAAAT-[69]-TAAAT-[186]-%ATTAG-[137]-aacaag-[4]-TAAAT-[89]-CTAAT-[132]-%ATTTA-
SERP1 NG1	Hs.151242	[-1996]-%ATTAG-[34]-%cttggt-[42]-%ATTAG-[14]-TAAAT-[20]-%ATTAG-[392]-%ATTAG-[48]-%ATTTA-[58]-%ATTTA-
ALB	Hs.184411	[-1983]-%ATTAG-[69]-%ATTTA-[34]-TAAAT-[3]-CTAAT-[18]-TAAAT-[52]-%ATTAG-[92]-CTAAT-[41]-%ATTTA-[82]-%cttggt-[24]-%cttggt-[139]-%ATTAG-[3]-%ATTAG-[179]-%ATTTA-[25]-CTAAT-[134]-%ATTTA-[6]-%ATTTA-
GC	Hs.198246	[-1919]-CTAAT-[44]-TAAAT-[43]-CTAAT-[110]-TAAAT-[44]-%ATTAG-[1]-%ATTTA-[24]-aacaag-[186]-TAAAT-[71]-TAAAT-[66]-%ATTAG-[24]-TAAAT-[197]-TAAAT-
ALDD B	Hs.234234	[-1924]-%ATTTA-[58]-aacaag-[4]-%ATTAG-[116]-%cttggt-[382]-%cttggt-[16]-TAAAT-[200]-TAAAT-[83]-%ATTTA-
APOA 2	Hs.237658	[-1693]-%ATTTA-[215]-CTAAT-[4]-%ATTAG-[341]-aacaag-[54]-%ATTTA-
TF	Hs.284176	[-1496]-%ATTTA-[238]-%ATTAG-[72]-%ATTAG-[22]-aacaag-[139]-%ATTTA-

Table 5 shows an example of occurrences of the association, “aacaag, HS\$GG_22 => HS\$GRH_03”, in liver. The genes involving the association are shown in the first column, and the UniGene ID is shown in the second one. The third one gives the detailed locations of known site homologs or putative regulatory sites in the promoter regions. For example, the fourth row in Table 6, the gene “AMBP” of UniGene cluster IDs of “Hs.76177” contains the association of “aacaag, HS\$GG_22 => HS\$GRH_03”. The “[-1999]-%cttggt-[227]-%ATTTA-[94]-%ATTTA-[6]-TAAAT-[135]-%ATTAG-[153]-TAAAT-” shows consequently the occurrence positions of the known site homologs and over-represented repetitive oligonucleotides within the associations, i.e., “HS\$GG_22/

CTAAT” or “HSSGRH_03/TAAAT” or “acaag”. The first number -1999 in bps is the location of the site “cttgt” from the transcription start site, while site “cttgt” is the reverse complement of site “acaag”. The symbol “%” means the site occurs at anti-sense strand. The distance of the first and second occurrences is 227 bps from transcriptional start site.

Table 6. Consensus of putative regulatory sites.

Prediction of Regulatory Sites							
Tissue Type	Putative Regulatory Sites	<i>Ms</i>	Occ	Exp	<i>sig</i>	Z-score	Consensus
Ovary (10 genes)	AAGAGG	10	26	8.06	3.08	6.32	MRGAGGM
	AGAGGC	8	24	5.70	4.60	7.59	
	GGAGGA	9	37	10.3	7.01	8.52	
	CGGAGG	10	41	14.5	4.75	6.98	
	CCAGGC	9	35	11.4	4.49	6.99	GCCAGSC
	CCAGCC	8	29	11.6	1.59	5.11	
	GCCAGC	8	15	3.81	1.62	5.73	
	ACCTCC	9	26	6.86	3.21	7.31	CASCTCC
	CACCTC	8	22	6.90	2.12	5.75	
	CAGCTC	9	24	6.24	3.25	7.12	
Liver (21 genes)	ACCTCA	20	40	16.9	2.57	5.60	ACCTCWCA
	CCTCTC	20	56	15.1	1.21	5.31	
	CTCTCA	19	36	15.9	1.65	5.03	
	CCTTCC	17	67	17.3	4.31	6.28	
	CCTGCA	17	39	12.3	5.67	7.61	
	CCTGCC	17	53	19.4	6.29	7.62	
	AGGCAG	17	57	26.6	3.34	5.88	CAGGCAGA
	CAGGCA	20	40	20.8	2.59	5.54	
GGCAGA	18	13	17.9	6.10	5.91		
Skeletal Muscle (35 genes)	CCTGGC	32	114	32.03	10.8	14.5	CYTGGCC
	CTTGGC	28	54	22.53	4.52	6.57	
	CTGGCC	30	99	25.48	10.6	14.6	
	CCAGAG	31	69	24.2	9.81	9.12	GMCAGAGA
	ACAGAG	30	88	42.1	5.06	7.07	
	GACAGA	31	82	32.7	9.14	8.61	
	CAGAGA	31	93	39.3	6.39	8.56	
	GGCTCA	32	107	34.6	9.06	12.28	GGCTCMG
	GGTCT	30	80	31.7	8.98	8.56	
GCTCAG	34	105	44.4	4.08	9.09		

The repetitive oligonucleotides associated with the known TF binding site are selected as putative regulatory sites because of their occurrence correlation to known signal of transcriptional regulation in a group of considered genes, i.e., these genes are differentially expressed in a specific tissue. As shown in Table 6, *Ms*, the matched upstreams, the number of gene upstreams which contain at least one pattern occurrence of the site; Occ,

the number of occurrences of the pattern among all upstream regions from genes which are higher expression in specific tissue; Exp, the expected number of occurrences; *sig*, significant coefficient calculated as defined in [15]; Z-score, calculated as defined before. For example, the sequence “AAGAGG” is located into 10 gene upstreams in the gene set of ovary. Twenty-six occurrences are found, while the expected occurrence is 8.06 in the 10 gene upstreams of length 2,000 bps. The *sig* value of “AAGAGG” in the liver is 3.08 and the Z-score is 6.32; “AGAGGC” are 4.60 and 7.59; “GGAGGA” are 7.01 and 8.52; “CGGAGG” are 4.75 and 6.98, respectively. The four over-represented repetitive oligonucleotides are extracted from the significant associations discovered in ovary. The consensus sequence “MRGAGGM” is generated by aligning these four sequences.

We further compute the D-value for each combination mined in each group of genes with higher expression in one tissue than in others to observe the dependence of a combination of regulatory sites in different tissues. As shown in Table 7, we select three tissues as example, ovary, liver, and skeletal muscle, to show the differential combinations (D-value exceeds 2.0, at least one support value greater than 0.4) of regulatory sites. The number of genes in each tissue is shown in the third row.

Table 7. The tissue-specific combinations of regulatory sites in three different tissues.

		Tissues (Ovary, Liver, Skeletal muscle)						
		Ovary	Liver	Skeletal muscle	Ovary	Liver	Skeletal muscle	
Tissues	Number of Genes	10	21	35	10	21	35	
	Tissue-specific Combination	χ^2 values			Support			D
Ovary	ggagcc, tgagca	(10.00)	1.64	0.44	(0.50)	0.00	0.17	3.47
	ggagaa, gcaaac	(6.67)	0.10	1.24	(0.50)	0.29	0.26	5.22
	cttggc, gaggcc	(10.00)	2.05	2.76	(0.50)	0.10	0.14	6.75
	GGTTA, gaggaa	(6.67)	0.04	0.00	(0.50)	0.29	0.29	4.41
	CTGTC, ccagtc	(4.44)	1.16	2.67	(0.80)	0.29	0.40	3.10
	ATTGG, gctgga	(4.29)	0.06	2.08	(0.50)	0.29	0.14	2.57
Liver	cagcac, gtctca	1.67	(8.03)	0.73	0.20	(0.43)	0.23	2.70
	agcctg, ccagca	(6.67)	(8.24)	0.97	0.00	(0.48)	0.29	3.93
	TAAAT, actaat	2.50	(4.49)	0.73	0.20	(0.48)	0.17	3.36
	GGGAAG, gaggaa	0.62	(8.24)	0.35	0.30	(0.48)	0.20	7.12
	CATTA, ctctca	0.10	(3.88)	0.24	0.10	(0.43)	0.20	2.57
Skeletal muscle	gacaga, gaggcc	1.67	0.40	(5.73)	0.10	0.14	(0.46)	3.21
	acatgg, cctggc	1.07	0.02	(10.01)	0.20	0.10	(0.46)	3.43
	GGCGG, cctgga	0.40	0.30	(4.61)	0.20	0.10	(0.43)	3.03
	GGCGG, gaggcc	0.40	1.54	(7.36)	0.30	0.00	(0.49)	7.99
	TATAA, agcctc	0.48	(8.24)	(7.20)	0.20	0.14	(0.54)	3.61

For example, in the fifth row the combination “ggagcc, tgagca” occurs in the gene upstreams of ovary, liver and skeletal muscle, where the Chi-square values are 10.0, 1.64, and 0.44, respectively. Chi-square values greater than 3.84 are shown with parentheses. Similarly, the support values are 0.5, 0, and 0.17. Support values, which are greater than

0.4, are shown with parentheses. The D-value is shown in the last column and the D-value of the combination “ggagcc, tgagca” is 3.47. Six combinations of “ggagcc, tgagca”, “ggagaa, ggcaac”, “cttggc, gaggcc”, “GGTTA, gaggaa”, “CTGTC, ccagtc” and “ATTGG, gctgga” are differential in ovary from considering the D-value of the six combinations greater than threshold 2.0. All their support values are also greater than 0.4, while in other categories the support values are much less than 0.4. Similarly, some combinations are also differential in other tissues.

4. DISCUSSION

To investigate the transcription regulation of human genes, in this study we attempt to find how combinations of known regulatory site homologs and over-represented repetitive oligonucleotides located within the promoter regions of differentially expressed gene groups which are found by statistical means in several considered tissues. Each gene promoter region is mapped to a “transaction” and known regulatory site homolog and over-represented repetitive oligonucleotides are mapped to items of a transaction. Data mining techniques are then applied to mine the associations. The enormous number of associations makes it extremely difficult to identify which are interesting and useful. The redundant rules are pruned and putative regulatory elements are obtained from the rest of the associations. Tissue-specific combinations of regulatory sites can be discovered statistically among the studied tissues and the significant values of the site combinations are provided.

Moreover, the proposed approach can also be potentially applied to other considered tissue sets, which can be different histologies, e.g., normal or cancer tissues. The approach can discover tissue-specific combinations of known site homologs and over-represented repetitive oligonucleotides between normal and cancer tissues to potentially investigate the differential gene regulation in transcription level. The thresholds of approaches like R-value, p-value, and D-value can also be adjusted when analyzing and studying different set of tissues.

Note that the co-occurrences of repetitive sequences and known site homologs reveal the repetitive elements to be putative regulatory elements because a set of transcription factor binding sites usually occur cooperatively. By considering the occurrence associations of known sites and repetitive sequences, the repetitive sequences can be viewed as putative regulatory signals correlated to the known site homologs. However, we find several associations that do not have any known Site homologs. The meanings and functionalities of these signals are interesting and necessary to be verified by biologists.

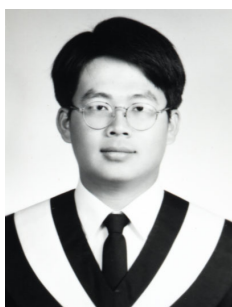
ACKNOWLEDGMENTS

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 89-2213-E-008-061. In addition, we would like to thank Prof. Cheng-Yen Kao at National Taiwan Univ. and Chi-Gong Tong at National Central Univ. for their valuable suggestions and comments.

REFERENCES

1. G. D. Schuler, M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tome, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannikulchai, A. Chu, C. Clee, S. Cowles, P. J. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, T. J. Hudson, and et al. "A gene map of the human genome," *Science*, Vol. 274, 1996, pp. 540-546.
2. M. S. Boguski, T. M. Lowe, and C. M. Tolstoshev, "dbEST – database for "expressed sequence tags," *Nat Genet*, Vol. 4, 1993, pp. 332-333.
3. J. S. Aaronson, B. Eckman, R. A. Blevins, J. A. Borkowski, J. Myerson, S. Imran, and K. O. Elliston, "Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data," *Genome Res*, Vol. 6, 1996, pp. 829-845.
4. M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, and et al., "Complementary DNA sequencing: expressed sequence tags and human genome project," *Science*, Vol. 252, 1991, pp. 1651-1656.
5. K. Okubo, N. Hori, R. Matoba, T. Niiyama, A. Fukushima, Y. Kojima, and K. Matsumura, "Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression," *Nat Genet*, Vol. 2, 1992, pp. 173-179.
6. S. Bortoluzzi, F. d'Alessi, C. Romualdi, and G. A. Danieli, "Differential expression of genes coding for ribosomal proteins in different human tissues," *Bioinformatics*, Vol. 17, 2001, pp. 1152-1157.
7. D. J. Stekel, Y. Git, and F. Falciani, "The comparison of gene expression from multiple cDNA libraries," *Genome Res*, Vol. 10, 2000, pp. 2055-2061.
8. S. Audic and J. M. Claverie, "The significance of digital gene expression profiles," *Genome Res*, Vol. 7, 1997, pp. 986-995.
9. E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhauser, M. Pruss, F. Schacherer, S. Thiele, and S. Urbach, "The TRANSFAC system on gene expression regulation," *Nucleic Acids Res*, Vol. 29, 2001, pp. 281-283.
10. A. Brazma, J. Vilo, E. Ukkonen, and K. Valtonen, "Data mining for regulatory elements in yeast genome," in *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, Vol. 5, 1997, pp. 65-74.
11. J. T. Horng, H. D. Huang, S. L. Huang, U. C. Yan, and Y. C. Chang, "Mining putative regulatory elements in promoter regions of *Saccharomyces cerevisiae*," *In Silico Biology*, Vol. 2, 2002, pp. 263-273.
12. J. T. Horng, H. D. Huang, M. H. Jin, L. C. Wu, and S. L. Huang, "The repetitive sequence database and mining putative regulatory elements in gene promoter regions," *Journal of Computational Biology*, Vol. 9, 2002, pp. 621-640.
13. H. D. Huang, H. L. Chang, T. S. Tsou, B. J. Liu, C. Y. Kao, and J. T. Horng, "A data mining method to predict transcriptional regulatory sites based on differentially expressed genes in human genome," in *Proceedings of Third IEEE Symposium on BioInformatics and BioEngineering*, 2003, pp. 297-304.
14. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "GenBank," *Nucleic Acids Res*, Vol. 31, 2003, pp. 23-27.

15. J. van Helden, B. Andre, and J. Collado-Vides, "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies," *Journal of Molecular Biology*, Vol. 281, 1998, pp. 827-842.
16. R. Srikant, Q. Vu, and R. Agrawal, "Mining generalized association rules," in *Proceeding of the 21st International Conference on Very Large Databases*, 1995, pp. 407-419.
17. B. Liu, W. Hsu, and Y. Ma, "Pruning and summarizing the discovered associations," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, pp. 125-134.



Hsien-Da Huang (黃憲達) was born in Taoyuan, Taiwan, in 1975. He received his B.S. degree in 1997 in Computer Science and Information Engineering in National Central University, Taiwan. He started his graduate studies on Bioinformatics in 1999, and received his Ph.D. degree in Institute of Computer Science and Information Engineering in National Central University, Taiwan. His current research interests are Bioinformatics, database systems, and data mining.



Huei-Lin Chang (張惠玲) was born in I-Lan, Taiwan, in 1978. She received Master degree in Computer Science and Engineering from Yuan Ze University, Taiwan in 2002. She is interested in the research of bioinformatics, database systems, data mining and wireless communication network.



Tsung-Shan Tsou (鄒宗山) is currently associate Professor of Biostatistics of the Graduate Institute of Statistics, National Central University, Taiwan. He received his B.S degree in Mathematics from National Taiwan University and got the M.S degree in Statistics from National Central University, Taiwan, in 1983 and 1987, respectively. He then spent 5 years in the Department of Biostatistics, School of Hygiene and Public Health, the Johns Hopkins University, Baltimore, USA, and received the Ph.D degree in 1992. His current research interests include robust statistical inferences methodology and the development of statistical tools for data mining technologies for bioinformatics.



Baw-Jhiune Liu (劉寶鈞) is a Professor in the Department of Computer Science and Information Engineering at Yuan Ze University in Taiwan since 1999. He received his B.S and M.S degrees in Electrical Engineering from National Cheng Kung University, Taiwan, in 1967 and 1969 respectively, and his Ph.D. degree in Electrical Engineering from National Taiwan University, Taiwan, in 1979. He worked for Telecommunication Labs. In Jungli, Taiwan from 1970 to 1973. He was Associate Professor in the Department of Computer Science and Information Engineering of National Taiwan University, Taiwan, from 1979 to 1983. He was a Professor in the Department of Computer Science and Information Engineering of National Central University, Taiwan, from 1983 to 1999. His current research interests include the development of data mining technologies for bioinformatics and the data models for web group learning.



Jorng-Tzong Horng (洪炯宗) was born in Nantou, Taiwan, on April 10, 1960. He received the Ph.D. degree in Computer Science and Information Engineering from National Taiwan University, Taipei, in April 1993. In 1993, he joined the Department of Computer Science and Information Engineering, National Central University, Jungli, Taiwan, where he became Professor in 2002. His current research interests include database systems, data mining, genetic algorithms, and bioinformatics.