A Computational Approach to Discover Differential Cooperation of Regulatory Sites in Functionally Related Genes in Yeast Genome

HSIEN-DA HUANG, JORNG-TZONG HORNG[1,2], CHIA-HUI CHANG[1],
TSUNG-SHAN TSOU[3], JING-YUE HONG[*] AND BAW-JHIUNE LIU[*]
Department of Biological Science and Technology
Institute of Bioinformatics
National Chiao Tung University
Hsinchu, 300 Taiwan
[1]*Department of Computer Science and Information Engineering*
[2]*Department of Life Science*
[3]*Institute of Statistics*
National Central University
Chungli, 320 Taiwan
E-mail: horng@db.csie.ncu.edu.tw
[*]*Department of Computer Science and Engineering*
Yuan Ze University
Chungli, 320 Taiwan

The availability of genome-wide gene expression data provides a unique set of genes from which to decipher the mechanisms underlying the common transcriptional response. A set of transcription factors which bind to target sites regulate the gene transcription cooperatively. This motivates us to discover the site associations of known transcription factor binding sites and certain repetitive elements. Those over-represented repetitive elements in the promoter regions of functionally related genes are predicted as putative regulatory elements. The study is to analyze how the differential site associations of the known regulatory sites and putative regulatory elements are distributed in the promoter regions of groups of functionally related genes. The functional-specific site associations are discovered by a statistical approach and the over-represented repetitive elements involving in the site associations are possible to be transcription factor binding sites. The site associations facilitate to predict functional-specific putative regulatory elements and to identify genes potentially co-regulated by the putative regulatory elements. Our proposed approach is applied to *Saccharomyces cerevisiae* and the promoter regions of yeast ORFs.

Keywords: regulatory site, transcription factor, yeast, data mining, promoter

1. INTRODUCTION

Identification of transcriptional regulatory elements within promoter regions is a topic of special interest for biologists since such elements govern the regulation of gene expression. Transcription factors (TFs), which are proteins, play a major role in gene

Received November 1, 2002; revised June 10, 2003; accepted August 5, 2003.
Communicated by Chuen-Tsai Sun.

regulation in eukaryotic organisms. The factors can bind to specific sites, specifically transcription factor binding sites or regulatory sites, in the promoter region of particular genes and interact with RNA polymerase and other factors to regulate the transcription of genes. Experimental regulatory profiles of known and unknown genes can be determined at a genomic scale thanks to the new technologies such as DNA microarray technology. De Risi *et al.* [1] have studied the diauxic shift in yeast and found several distinct gene groups by clustering the gene expression profiles. De Risi *et al.* also show the presence of several regulatory sites in promoter regions of the respective genes. Helden *et al.* [2] also studies the dataset constructed by De Risi *et al.* and systematically searched the promoter region of potentially co-regulated genes for over-represented oligonucleotides which called be transcription factor binding sites and involve the gene regulation. A systematic analysis of over-represented sequence patterns in clusters of promoter regions obtained by clustering the diauxic shift expression profiles has been done by Brāzma *et al.* [3]. It was shown that over-represented pattern occurrence in promoter regions for genes from expression profile clusters of at least 25 genes cannot be explained by statistical chance.

Many experimental identifying transcription regulatory sites have been collected in TRANSFAC [4], which is the most complete and well maintained database on transcription factors, their genomic binding sites and DNA-binding profiles [4]. Notably, consensus patterns or nucleotide distribution matrices can be used to describe transcription factor binding sites. While describing binding sites, Brāzma *et al.* [5] stated, "The matrix representation is generally considered to be the best available means for representing the consensus. However, at present, most consensus descriptions are unreliable in the sense that they tend to give many false positives when compared against the genome sequences of even modest length". Therefore, this study describes the binding sites using consensus patterns. Brāzma *et al.* [5] developed a general software tool for finding and analyzing the site associations of transcription factor binding sites that occur frequently in gene upstream regions in the yeast genome. In addition to analyzing the association rules of regulatory sites, their work focused on promoter and random regions, in which their ratio appears. Their tool can find all the site associations satisfying the given parameters with respect to the given set of gene promoter regions, its counterset, and the chosen set of sites. Previous research in Horng *et al.* [6, 7] also investigated the site associations of regulatory sites and over-represented repeats in yeast genome by a data mining approach, and some significant oligonucleotides are predicted as putative sites based on the their significant correlation to known sites.

Composite regulatory elements provided in COMPEL [8] contains two closely situated binding sites for distinct transcription factors and represents minimal functional units providing combinatorial transcription regulation [8]. They address the fact that the complex differential expression of genes in higher organisms is achieved through combinatorial regulation of transcription by specific combination of transcription factors binding to their target sites in the regulatory regions of these genes. Their database emphasizes the key role of specific protein-protein interactions for gene regulation in a particular cellular content [8]. In comparison with our proposed approach, a data mining approach is used to computationally discover the site associations of target sites which are more than two. We investigate the site associations of more than two target sites of regulatory elements in contrast to the two targets sites in COMPEL.

The RSDB [7] database contains repetitive elements, which are classified into exact, tandem, and similar elements. We observe that the transcription factor binding sites in TRANSFAC have the property of repetitiveness. This motivates as to discover the site associations of known transcription factor binding sites in TRANSFAC and certain repetitive elements. Those over-represented repetitive elements may or may not be putative regulatory elements. The repetitive sequences in our experiments include direct and inverted repetitive sequences whose length is between 5 and 25 nucleotides.

To deal with such a large amount of data, data mining plays a prominent role in knowledge extraction. The enormous number of sequenced genomes, gene identification data, gene expression experimental profiles, and genes categorized into functional classes allows the use of computational techniques to investigate transcriptional regulatory elements in the gene promoter regions and decipher the mechanisms of gene transcriptional regulation. Frequently used data mining approaches include association rules, statistics, neural network, clustering, classification, and genetic algorithms, etc. Agrawal *et al.* [9] introduced the problem of mining association rules over basket data. The data mining techniques might mine an enormous number of associations. Such a large number of associations makes it extremely difficult to identify those which useful or interesting. Chi-square testing is one of the approaches to remove insignificant ones by testing the occurrence correlation of the two events in a association [10].

This study initially identifies the site associations of known regulatory sites extracted from TRANSFAC [4] and over-represented repetitive oligonucleotides retrieved statistically from RSDB [7] in the promoter regions of a particular set of selected genes. Mining association rules are then applied to mine the associations from the combinations of over-represented repetitive elements and known regulatory sites. The site associations found in each functional gene group are then statistically analyzed in all other groups. A Chi-square test is applied to determine the dependence of the sites of the site associations; the R-value of each combination is computed among groups of genes to find its differential occurrences in each group of functionally related genes. Those target sites in highly dependent site associations with large R-values, i.e., small p-values, in a functional gene group are candidates of putative functional-specific regulatory sites in the group because of their specificities in that functional gene group.

2. OBSERVATION

We observe the repetitive property of yeast transcription factor (TF) binding sites in TRANSAC [4] by computationally locating the sites in the yeast genomic sequences and gene upstreams. The number of occurrences of yeast TF binding sites in TRANSAC are partially shown in Table 1. For instance, the site “GATAA” with accession number “[R00494]” occurs 39,395 times in the whole yeast genome, 12,401 times in the gene upstreams, and 9,982 times in the coding regions of genes. The site can be located in the coding regions or in the upstreams of 5,324 genes. The expectation of occurrence in upstreams is 12333.48 times, while in coding regions it is 9676.35 times. In comparison with the other oligonucleotides occurring in the gene upstreams, we compute the occurrences of the oligonucleotides in the upstreams, coding regions, and the whole genome as shown in Table 2. For example, the repetitive oligonucleotide “ACCCTA” given in the

Table 1. The occurrences of partial TF binding sites for yeast genome in TRANSAC [4].

Sites in TRANSFAC	Accession Number	Site ID in TRANSFAC	Amount of occurrences (times) in						No. of Genes Related
			Genome	Background probability	Up-streams	Exp.	Coding Regions	Exp.	
GGGG	[R00496]	Y$GAL1_11	29,694	$2.44*10^{-3}$	8,892	9296.37	8,417	7293.55	4,340
CCGA	[R00256]	Y$CYC1_03	38,751	$3.19*10^{-3}$	12,121	12131.86	9,798	9518.17	5,239
GAGGA	[R00492]	Y$GAL1_07	24,972	$2.05*10^{-3}$	7,859	7818.04	6,066	6133.72	4,127
GATAA	[R00494]	Y$GAL1_09	39,395	$3.24*10^{-3}$	12,401	12333.48	9,982	9676.35	5,324
AGCCT	[R00495]	Y$GAL1_10	15,199	$1.25*10^{-3}$	4,785	4758.38	3,994	3733.24	3,355
ATATAA	[R00497]	Y$GAL1_12	18,120	$1.49*10^{-3}$	4,853	5672.87	6,437	4450.70	4,216
GAGTCA	[R00645]	Y$HIS3_02	3,911	$3.22*10^{-4}$	1,243	1224.43	962	960.63	1,000
CAGTCA	[R00656]	Y$HIS4_13	4,472	$3.68*10^{-4}$	1,407	1400.06	1,116	1098.43	1,169
AAGTCA	[R00831]	Y$ILV1_03	7,945	$6.54*10^{-4}$	2,605	2487.36	1,848	1951.48	1,846
GATGACC	[R00260]	Y$CYC1_07	1,148	$9.44*10^{-5}$	367	359.41	269	281.98	302
ATGAAACA	[R01838]	Y$STE2_04	1,054	$8.67*10^{-5}$	339	329.98	261	258.89	322
ATGAAACC	[R01842]	Y$MFA2_01	621	$5.11*10^{-5}$	211	194.42	134	152.53	153
ATGTAAAT	[R01362]	Y$STE2_02	963	$7.92*10^{-5}$	278	301.49	329	236.54	376
AAGTACAT	[R01361]	Y$STE2_01	701	$5.77*10^{-5}$	220	219.46	197	172.18	236
ATGACTAAT	[R02023]	Y$TRP4_02	93	$7.65*10^{-6}$	28	29.12	35	22.84	42
AGCCGCCGA	[R02319]	Y$CAR1_01	41	$3.37*10^{-6}$	6	12.84	26	10.07	36

(Abbreviation: exp. denotes the expectation amount of site occurrence times in the upstreams or coding regions.)

Table 2. The occurrence amount of oligonucleotides in yeast genome.

Repetitive Oligonucleotide	Amount of occurrences (times) in						No. of Genes Related
	Genome	Background probability	Up-streams	Exp.	Coding Regions	Exp.	
ACCCTA	2,724	$2.24*10^{-4}$	822	853.51	793	669.63	834
ACCCTC	2,917	$2.40*10^{-4}$	881	913.98	795	717.07	851
AGTACT	3,073	$2.53*10^{-4}$	933	962.86	879	755.42	973
AGTAGA	6,673	$5.49*10^{-4}$	1,970	2090.85	1,798	1640.40	1,846
AGTAGC	4,912	$4.04*10^{-4}$	1,545	1539.08	1,299	1207.50	1,377
GATACC	4,829	$3.98*10^{-4}$	1,638	1513.07	1,005	1187.09	1,054
GATAGA	7,030	$5.79*10^{-4}$	2,163	2202.71	1,807	1728.16	1,824
TGGTAA	10,513	$8.66*10^{-4}$	3,493	3294.04	2,214	2584.37	2,129
AGAGTTA	1,859	$1.53*10^{-4}$	610	582.48	441	456.99	485
CAATCAG	1,358	$1.12*10^{-4}$	445	425.50	320	333.83	355
CGTCTCC	592	$4.87*10^{-5}$	199	185.49	148	145.53	169
CGTCTGA	652	$5.37*10^{-5}$	196	204.29	165	160.28	190
ACAAACTC	514	$4.23*10^{-5}$	175	161.05	112	126.35	133
CACAGAAGA	164	$1.35*10^{-5}$	57	51.39	39	40.32	46
AGAGTGG	983	$8.09*10^{-5}$	310	308.00	271	241.65	311

first row occurs 2,724 times in the whole yeast genome with probability 2.24 * 10^{-4} (The size of yeast genome is 12,146,300 bps). The expectation of occurrence times in upstreams is 853.51 times, and in coding regions it is 669.63.

From Tables 1 and 2, we find both known TF binding sites in TRANSFAC and repetitive elements have the property of repetitiveness. Therefore, the occurrences of repetitive sequences correlating to the occurrences of known sites can be investigated as putative binding sites, tissue-specific regulatory sites, functional-specific binding sites, or other unknown regulatory signals. After statistically analyzing and data mining, the discovered association of site occurrences also reveals that the sites potentially regulate the transcription of a particular set of genes.

3. METHODS

The proposed approach is given as follows. We first preprocess the target sites and gene promoter regions to find the site associations of known sites and over-represented oligonucleotides located in the promoter regions of the groups of functionally related genes. Next, a mining association rule algorithm [9] is applied to mine the association rules by combining the known sites and over-represented repeats. A Chi-square test is then used to select certain interesting and significant rules. The R-value of each site combination is computed to investigate the differences of transcriptional regulation in different functional gene categories. Finally, the over-represented repeats within the significant and differential site associations, which are mapped to the items in the association rules, are selected as putative regulatory sites [6].

Materials

Before analyzing of the associations of known site homologs and over-represented repetitive sequences located in the upstream regions, we collect the sequence of yeast genome and the gene annotations from NCBI[1]. 6,350 yeast genes and ORFs are documented in MIPS [11], and 3,529 genes are classified into at least one functional category. The experimental identifying transcription factor binding sites can be obtained from TRANSFAC [4]. The TRANSFAC database (professional 5.4) contains 11,537 site sequences of which the number of yeast sites is 285. Most sites are also consensus patterns. The data in TRANSFAC has the following features. A transcription factor binding site accession number may have 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 within larger ones.

During the detection of over-represented oligonucleotides, the occurrences of the oligonucleotides in yeast genome are necessary. Repetitive sequences with lengths from 10 to 25 bps of the yeast genome can be obtained from the repetitive sequence database (RSDB) [7]. For oligonucleotide lengths from 4 to 9 bps, we proposed an algorithm to construct the yeast genome sequence into a special computational data structure, i.e., Suffix-array [12], to reduce the algorithmic complexity when searching an oligonucleo-

[1] http://www.ncbi.nlm.nih.gov/.

tide in a yeast genome sequence. Accordingly, the occurrences of a query oligonucleotide are returned efficiently by querying the suffix-array of the under consideration genome.

Typically, the lengths of the query oligonucleotides in the application of regulatory site prediction do not exceed 25 bps. We construct the suffix-array and support the querying of occurrences of oligonucleotides whose length is from 4 to 25 bps. When performing the oligonucleotide analysis to discover over-represented repeats in the upstream regions of genes, the frequencies of occurrence of all possible oligonucleotides whose length is from 4 to 25 bps can be efficiently returned by querying in the suffix-array. Due to the frequent update of the genome assembly sequences, the method is designed to efficiently reconstruct the index of whole genome sequences.

Preprocessing and Mapping

First, the transcription factor binding sites categorized in yeast from TRANSFAC and repetitive oligonucleotides in RSDB and the suffix-array are prepared. For each group of functionally related genes, all of the known regulatory sites in yeast are directly located the promoter region from – 1 to – 800 bps (+ 1 denotes the gene translational start site), and the repetitive oligonucleotides are also located. The occurrences of each known site and repeats are calculated and provided for statistical analysis. The occurrence combination of the known sites and repeats within each promoter region are stored for data mining.

Statistical Analysis of Over-Represented Oligonucleotides

To detect the over-represented (OR) 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 sequence [2]. 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 statistically 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 the whole yeast genome, where the occurrence frequencies of oligonucleotides are obtained from the suffix-array. 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 [2]. A specific expected frequency has thus to be used for each oligonucleotide sequence. Helden *et al.* proposed a statistical method to estimate the probability of observing exactly *n* occurrences of the oligonucleotide *b* within promoter regions of a gene family by the binomial formula. 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 sequences of Yeast are obtained by applying the statistical method in [2]. The oligonucleotides, which have significant values exceeding the threshold, are selected as significant over-represented ones.

We classify the statistics according to several groups of genes. These datasets make ideal datasets to use for data mining. The function catalogues are collected from MIPS [11]. Then a specific expected frequency is used for each repetitive oligo-mers to determine the statistical significance.

$$T = 2 \times S \times (L_i - w + 1) \tag{1}$$

$$P(occ\{b\} = n) = \frac{T!}{(T-n)! \times n!} \times (F_e\{b\})^n \times (1 - F_e\{b\})^{(T-n)} \tag{2}$$

$$P(occ\{b\} \geq n) = \sum_{j=n}^{T} P(occ\{b\} = j) = 1 - \sum_{j=0}^{n-1} P(occ\{b\} = j) \tag{3}$$

where $F_e\{b\}$ 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 ith sequence in the set; $P(occ\{b\} = n)$ is the probability of observing exactly n occurrences of the oligomer b; $P(occ\{b\} \geqq n)$ is the probability to observe n or more occurrences of the oligomer b.

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

D is the distinct number of oligomers; N_{pal} is the palindromic oligomers. Lastly we define a significance coefficient

$$sig = -\log_{10} [P(occ\{b\} \geq n) \times D] \tag{5}$$

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

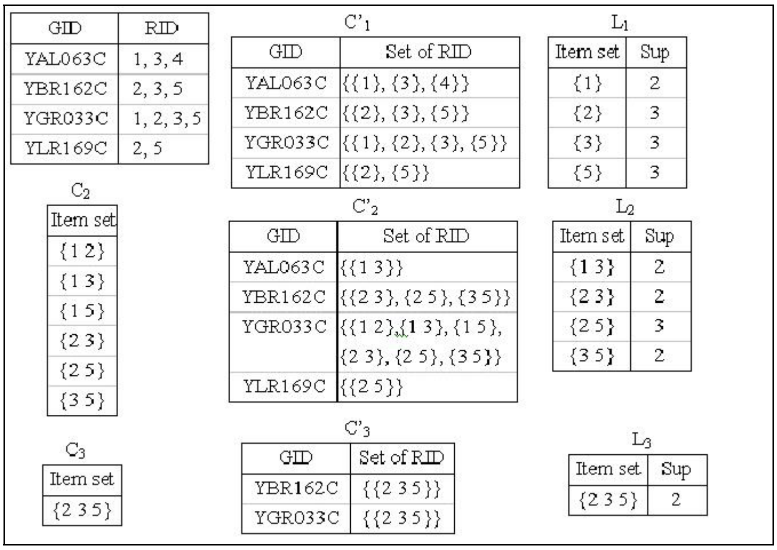
Mining Site Occurrence Associations

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 with transactions, where each transaction consists of a set of items. An association rule is an expression such as $A => B$, where A and B are the sets of items. The related mining 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 $A => B$ used here is the percentage of transactions that contain both A and B.

The formal statement of the problem is described below. Let $S = \{s_1, s_2, \ldots, s_m\}$ be a set of known sites of yeast in TRANSFAC and $R = \{r_1, r_2, \ldots, r_n\}$ be a set of over-represented repetitive sequences in yeast from RSDB. The union of the sets S and R is called 'item set'. Let $G = \{g_1, g_2, \ldots, g_m\}$ 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 sites and over-represented repeats, 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 $A => B$, where $A \subset I$, $B \subset I$, and $A \cap B = 0$. The rule *A* => *B* holds in the set of promoter regions *D* with *confidence conf* if *c*% of transactions in *D* contains both *A* and *B*. The rule *A* => *B* has *support sup* in the repetitive sequence set *D* if *s*% of promoter regions in *D* contain $A \cup B$. The association rules are generated if the rule has a higher support and confidence than specified by the user. The a priori algorithm [9] is then implemented to mine association rules.

Fig. 1 presents an example of the mapping between the gene promoter regions and regulatory sites, i.e., known site homologs and over-represented oligonucleotides. GID denotes the identities of gene upstreams, and RID denotes the identities of the regulatory sites, i.e., known site homologs or OR oligonucleotides. For example, YAL063C gene upstream contains the regulatory sites RID{1, 3, 4}. L_i denotes the phase of discovering combinations of length *i*. The combination of {2, 3, 5} in L_3 with support 2 means two genes of YBR162C and YGR033C contain the site associations.



GID	RID
YAL063C	1, 3, 4
YBR162C	2, 3, 5
YGR033C	1, 2, 3, 5
YLR169C	2, 5

C'1

GID	Set of RID
YAL063C	{{1}, {3}, {4}}
YBR162C	{{2}, {3}, {5}}
YGR033C	{{1}, {2}, {3}, {5}}
YLR169C	{{2}, {5}}

L1

Item set	Sup
{1}	2
{2}	3
{3}	3
{5}	3

C2

Item set
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

C'2

GID	Set of RID
YAL063C	{{1 3}}
YBR162C	{{2 3}, {2 5}, {3 5}}
YGR033C	{{1 2},{1 3}, {1 5}, {2 3}, {2 5}, {3 5}}
YLR169C	{{2 5}}

L2

Item set	Sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C3

Item set
{2 3 5}

C'3

GID	Set of RID
YBR162C	{{2 3 5}}
YGR033C	{{2 3 5}}

L3

Item set	Sup
{2 3 5}	2

Fig. 1. An illustrative example of a mining association of regulatory sites.

Filtering Insignificant Site Associations

In the site co-occurrence detection step, the site combinations co-occurring in the upstream regions are detected. In order to filter insignificant site combinations, two statistics, Chi-sequence tests and cumulative hypergeometric distribution, are incorporated. The basic idea is that the sites in the left-hand-side part and right-hand-side part of a site combination may emerge independently in the upstream regions in a group. Note that a site combination is divided into left-hand-side and right-hand-side parts and is denoted as the example of “aaatat, ttgaa => gcggag”. To filter insignificant ones, a Chi-square test is performed to investigate each site combination to test the hypothesis that the occurrence of sites in the left-hand-side part of the site combination is independent of the site in the right-hand-side part. In rejecting the hypothesis, the site combination is considered to be a significant site combination. This means that all sites in the left-hand-side part occur significantly concurrently in the upstream regions with the site in the right-hand-side part of the site combination, if the chi-square value exceeds 3.84.

In statistics, the Chi-square test (χ^2) is widely used for testing independence and/or correlation and is applied to discover significant associations rules in [10]. A huge number of combinations is found in each group of functionally related genes and then a Chi-square test is used to investigate the dependence of the sites of each combination in all functional gene groups. However, any combinations may have very different site dependence among the genes in the group when the Chi-square values exceed a threshold of 3.84 with degree of freedom 2 and $\alpha = 0.05$. Let f_0 be an observed frequency and f an expected frequency, Chi-square is used to test the significance of the deviation from the expected values. The value of χ^2 is defined as

$$x^2 = \sum \frac{(f_0 - f)^2}{f} \tag{6}$$

In Fig. 2, we would like present an example to show how to test the correlation of the sites in the combination “aaatat, ttgaa”. The two sequences, “aaatat” and ”ttgaa”, occur concurrently in 23 of 35 genes in the category “Drug Transporters”. Six genes cannot be found in any of the two sequences. Five genes contain the sequence “ttgaa”, but not “aaatat”. The four conditions are constructed as a 2 by 2 contingency table shown in Fig. 2.

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

Fig. 2. A contingency table to show the genes containing sites “aaatat” and “ttgaa”.

If the correlation of the two sites is independent, we expect the total number of “aaatat” site, i.e., 24, will be divided into proportions of 80% (28/35) and 20% (7/35). A χ^2 value of 0 implies the sites are statistically independent. If it is higher than 3.84 is at the 95% significance level with 1 degree of freedom, we reject the hypothesis of independence of site occurrences, and say that the sites occurrences are correlated.

$$X^2 = \frac{(24*0.8-23)^2}{24*0.8} + \frac{(24*0.2-1)^2}{24*0.2} + \frac{(11*0.8-5)^2}{11*0.8} + \frac{(11*0.2-6)^2}{11*0.2} = 11.965$$

Detecting Differential Site associations

In order to investigate the occurrence of the same site combination mined in different functional categories, we propose the R-statistic to compute the hypothesis that the site combination occurrence in each category is consistent with the others. If rejecting the hypothesis, i.e., the R-value is higher than a specified threshold 2.0 then the biological meaning is the site combination occurs differentially in particular functional category. Computing R-statistic can extract the site associations whose occurrence varied most across different functional categories. The statistic is denoted as R_j for each combination j given by Eq. (7), where m is the number of functional categories, $X_{i,j}$ is the occurrences of the combination j in the functional category i, and N_i is the total number of genes, i.e., ORFs, in the ith functional category. The frequency f_j of the combination j in all of the functional categories is given by Eq. (8).

$$R_j = \sum_{i=1}^{m} X_{i,j} \ln\left(\frac{X_{i,j}}{N_i f_j}\right) \tag{7}$$

$$f_j = \frac{\sum_{i=1}^{m} x_{i,j}}{\sum_{i=1}^{m} N_i} \tag{8}$$

For instance, the number of ORFs in the numerics of MIPS functional categories of 0104, 0201, 0316, 0510, and 0722 are 31, 34, 90, 36, and 43, respectively. The occurrences of the combination, “tataca, ttgaaa”, is 16, 4, 26, 12, and 13, respectively. The R-value in the example is 4.52 as follows. The greater the R value computed from different functionally gene groups, the more differential the combination among the select groups is.

$$f_j = \frac{16+4+26+12+13}{31+34+90+36+43} \approx 0.3$$

4. RESULTS

Table 3 shows detailed information on transcription factor binding sites in TRANSFAC, and over-represented oligonucleotides found in different functional categories in the yeast genome. For example, the first row in Table 3 indicates about 157 over-represented repeats are selected after applying statistical analysis in the functional category of “amino-acid transport”. Also, 33 known sites in TRASFAC can be located in the gene promoter regions in this category. We then mine the associations from the site associations of these over-represented repeats and known site homologs.

Table 4 shows the associations mined by our proposed approach in each group of functionally related genes. The minimum support and confidence are set to 40%. As given in Table 4, 110 associations are discovered in 23 promoter regions in the function category of “Amino-acid transport”, and 121 associations are discovered in 23 promoter regions in the function category of “Amino-acid transport“. After pruning by chi-square testing, 11 significant associations are found.

Fig. 3 shows an example of the occurrence of the association, e.g., “RAF, GAL4, DBF-A => gaaata”, in the functional category of “Purine and pyrimidine transporters”. The gene YPL134C with the annotation “mitochondrial 2-oxodicarboxylate transport

Table 3. The number of known site homologs and OR oligonucleotides in the gene upstreams of gene functional categories.

MIPS functional category	MIPS in numeric	Amount	
		OR oligonucleotides	**Known sites homologs**
Amino-acid transport	01.01.07	157	33
Deoxyribonucleotide metabolism	01.03.07	134	29
Polynucleotide degradation	01.03.16	156	33
Nucleotide transport	01.03.19	70	36
Lipid and fatty-acid transport	01.06.13	82	37
Other proteolytic degradation	06.13.07	64	29
Anion transporters (Cl, SO4, PO4, etc.)	07.04.07	85	36
Homeostasis of protons	13.01.02	175	45

Table 4. The associations of known sites and repeats mined in each functional category.

MIPS functional category	Amount			
	ORFs	**Site associations** (before pruning)	**Filter by chi-square value**	**Significant site associations**
Amino-acid transport	23	121	110	11
Deoxyribonucleotide metabolism	12	156	143	13
Polynucleotide degradation	22	90	78	12
Nucleotide transport	16	59	54	5
Lipid and fatty-acid transport	18	70	64	6
Other proteolytic degradation	13	73	64	9
Anion transporters (Cl, SO4, PO4, etc.)	17	123	111	12
Homeostasis of protons	32	84	65	19

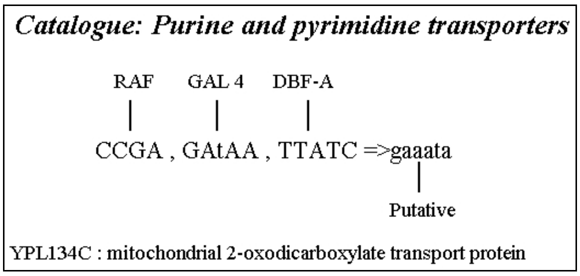

Catalogue: Purine and pyrimidine transporters
RAF
GAL 4
DBF-A
CCGA , GAtAA , TTATC =>gaaata
Putative
YPL134C : mitochondrial 2-oxodicarboxylate transport protein

Fig. 3. An example of prediction of putative regulatory sites.

protein" is categorized as "Purine and pyrimidine transporters" in MIPS [11]. The association rule shown in Fig. 3 consists of three known sites, "CCGA/RAF", "GAtAA/GAL4", and "TTATC/DBF-4", and one OR oligonucleotide, "gaaata". Note that the

OR oligonucleotide in the site association are predicted as putative regulatory sites in the functional category of “Purine and pyrimidine transporters”.

Several interesting and significant association rules in each functional category are given in Table 5. Column one in Table 5 is the functional categories in MIPS; column two is the associations containing known regulatory sites in *uppercase* and over-represented oligonucleotides in *lowercase*; column three is the confidence value of the association; column four is the support value; column five is the Chi-square value; and column six is similar to the second column but using the TF binding site names instead. For instance, the association of “TTATC => cgccg” is significant, which was discovered in “Amino-acid transport”; the support value is 0.522, the confidence value is 0.632, and the χ^2 value is 5.282. “TTATC” are homologous to the known sites Y$ARS1_05 and “cgccg” is a significant over-represented oligonucleotide.

Table 5. Significant associations mined in each of the MIPS functional categories.

MIPS functional category	**Site associations**	**Conf.**	**Sup.**	χ^2	**Site associations** (names of homologous known site with TRANSFAC ID)
Amino-acid transport	TTATC=>cgccg	0.632	0.522	5.282	Y$ARS1_05=>cgccg
	ATATAA=>TTATC	0.938	0.652	4.542	Y$GAL1_12=>Y$ARS1_05
	ATATAA=>GATAA	0.938	0.652	7.413	Y$GAL1_12=>Y$GAL1_09
	GGGG=>aagcg	0.789	0.652	4.542	Y$GAL1_11=>aagcg
	ATATAA=>cggcaa	0.625	0.435	4.537	Y$GAL1_12=>cggcaa
	cgtgc=>gcgcc	0.786	0.478	4.707	cgtgc=>gcgcc
	cgtgc=>gccgc	0.786	0.478	7.078	cgtgc=>gccgc
Deoxyribonucleotide metabolism	CATCC=>TATAAA	0.857	0.5	5.182	Y$ENO2_03=>Y$CUP1_07
	GATAA=>aacgc	0.9	0.75	7.2	Y$GAL1_09=>aacgc
	cgcac=>cgtcc	1	0.5	6	cgcac=>cgtcc
	aaacg=>cgcgta	0.667	0.5	4	aaacg=>cgcgta
	aaacg=>acccg	0.778	0.583	5.6	aaacg=>acccg
Polynucleotide degradation	ATATAA=>aatatta	0.667	0.455	5.238	Y$GAL1_12=>aatatta
	atatata=>atattaa	0.588	0.455	5.392	atatata=>atattaa
Nucleotide transport	TATAAA=>TTATC	0.933	0.875	7.467	Y$CUP1_07=>Y$ARS1_05
	GGGG=>aggcg	0.875	0.438	6.349	Y$GAL1_11=>aggcg
	ATATAA=>cgcgc	0.583	0.438	4.148	Y$GAL1_12=>cgcgc
Lipid and fatty-acid transport	CATCC=>GATAA	1	0.611	5.657	Y$ENO2_03=>Y$GAL1_09
	ATATAA=>GATAA	1	0.778	12.6	Y$GAL1_12=>Y$GAL1_09
Other proteolytic degradation	GAGGA=>TTATC	1	0.615	6.24	Y$GAL1_07=>Y$ARS1_05
	TTATC=>cgcga	0.7	0.538	4.55	Y$ARS1_05=>cgcga
	atcgc=>cacgc	0.778	0.538	6.741	atcgc=>cacgc
	cacgc=>cgatc	0.857	0.462	6.198	cacgc=>cgatc
Anion transporters (Cl, SO4, PO4, etc.)	CATCC=>GATAA	1	0.611	5.657	Y$ENO2_03=>Y$GAL1_09
	ATATAA=>GATAA	1	0.778	12.6	Y$GAL1_12=>Y$GAL1_09
	ATATAA=>GGGG	0.75	0.529	4.408	Y$GAL1_12=>Y$GAL1_11
	agggc=>tcgca	0.917	0.647	5.236	agggc=>tcgca
	agggc=>cacac	0.75	0.529	7.969	agggc=>cacac
Homeostasis of protons	GAGGA=>GATAA	1	0.562	4.256	Y$GAL1_07=>Y$GAL1_09
	ATATAA=>attataa	0.591	0.406	6.732	Y$GAL1_12=>attataa
	ATATAA=>GGGG	0.773	0.531	6.555	Y$GAL1_12=>Y$GAL1_11
	TTATC=>atatata	0.538	0.438	5.744	Y$ARS1_05=>atatata
	TTATC=>atataat	0.538	0.438	5.744	Y$ARS1_05=>atataat
	attata=>taaata	0.8	0.5	4.885	attata=>taaata
	ataatat=>taaata	1	0.438	13.037	ataatat=>taaata
	ataaata=>ataata	0.929	0.406	5.42	ataaata=>ataata

The detailed positions of known sites and putative regulatory sites in the association "Y$ARS1_05 => cgccg" in a set of ORFs are shown in Table 8. The first column gives the ORFs of yeast in the set and the second shows the detailed positions of known sites and putative regulatory sites in each ORF. For example, the first row in Table 6, "YBR069C" is the ORF name, and "[-587]-TTATC-[327]-%GATAA-[77]-%cggcg-[0] -%cggcg-[71]-TTATC-" is the composition of known and putative regulatory sites. The first number "[-587]" denotes the offset of the site "TTATC" from the start position of the coding region either in direct or reverse strain, the symbol "%" denotes the site appearing on the other strand, and the distance between "TTATC" and "GATAA" is "[327]" bps, and so on.

Table 6. The occurrence of known and putative regulatory sites in the association (Y$ARS1_05 => cgccg).

ORFs	Occurrences of sites in "Y$ARS1_05 => cgccg"
YBR069C	[-587]-TTATC-[327]-%GATAA-[77]-%cggcg-[0]-%cggcg-[71]-TTATC-
YBR132C	[-482]-%GATAA-[140]-%cggcg-
YCL025C	[-351]-TTATC-[27]-%GATAA-[139]-TTATC-[124]-%cggcg-
YPL265W	[-541]-%cggcg-[25]-%cggcg-[76]-TTATC-[176]-TTATC-[13]-TTATC-[117]-%GATAA-
YPL274W	[-598]-%GATAA-[163]-TTATC-[80]-%cggcg-
YNL270C	[-360]-TTATC-[36]-cgccg-
YLR375W	[-416]-cgccg-[1]-cgccg-[94]-TTATC-[138]-cgccg-
YNL268W	[-116]-%cggcg-[79]-%GATAA-
YHL036W	[-580]-%cggcg-[24]-TTATC-[37]-TTATC-[129]-%GATAA-[96]-%GATAA-[189]-TTATC-

Table 7. The regulatory families and their regulatory property [2].

Family	Genes	Common regulatory property	Reference
NIT	DAL5, DAL80, GAP1, MEP1, MEP2, MEP3, PUT4	Repressed when good nitrogen sources (glutamine glutamate, ammonia) are present in the medium	Magasanik (1992)
PHO	PHO5, PHO11, PHO8, PHO84, PHO81	Repressed by P_i	Oshima *et al.* (1996)
MET	MET3, MET2 MET14 MET6 SAM1 SAM2 MET1 MET30 MUP3	Repressed by methionine	Hinnebusch (1992), Blaiseau *et al.* (1997)

We also apply our proposed approach to the previously characterized regulatory families in [13, 14], which are also investigated in [2]. For each family in Table 7, we extracted the 600 bps upstream sequences in the genes, and used our approach to discover the associations of known site homologs and over-represented oligonucleotides. Table 8 shows ms (matching sequences, i.e., the number of genes from the family which contain at lease one occurrence of the site), occ (the number of occurrences of the site among all promoter regions from the family), exp (the expected number of occurrences),

Table 8. Alignment of OR oligonucleotides.

Gene Family	Putative regulatory elements	Ms	Occ	Exp	*Sig*	Consensus	Site previously characterized	
							Consensus	**Bound factors**
NIT	ATATAA	6	10	12.05	-0.20	AKATAAGA	GATAAG	Gln3p, Nillp, Gzf3p, Uga43p (Zn finger)
	GATAAG	6	25	4.04	**8.40**			
	ATAAGA	6	19	6.34	**1.13**			
	GGCAC	5	10	6.89	-0.91	GGCACA	--	--
	GCACA	5	11	9.46	-0.26			
	TGTGC	5	11	9.46	-0.29	TGTGTT	--	--
	GTGCC	5	10	6.89	-0.92			
PHO	CGCAC	4	9	3.015	-0.43	CGCACG	GCACGTGGG	Pho4p (bHLH)
	CGCACG	4	5	0.52	**0.37**			
	ACGTATA	4	6	0.72	**0.07**	ACGTATATA	--	--
	ACGTATATA	4	4	0.13	-0.14			
MET	TCACGT	8	17	2.71	**5.00**	TCACGTGA	TCACGTG	Cbflp-Met4p-Met28p complex (Zn finger)
	TCACG	8	21	8.80	**0.77**			
	CACGTG	8	11	0.83	**5.51**			
	CACGT	8	23	8.83	**1.59**			
	ACGTGA	8	17	2.71	**5.00**			
	CGTGA	8	21	8.80	**0.77**			

and *sig* (significant index, calculated as in [2]). For instance, in the NIT family, the significant value, *sig*, of the sites “ATATAA”, “CATAAG”, and “ATAAGA” are – 0.20, 8.40, and 1.13, respectively. The consensus sequence “AKATAAGA” aligned from the putative regulatory sites is similar to the previously characterized consensus “GATAAG” [13]. Similarly, in the PHO family, the consensus “CGCACG” is also aligned from putative sites “CGCAC” and “CGCACG”, and is similar to the consensus “GCACGTGGG” characterized in [14].

We further compute the R-value for each combination mined in each functional gene group to observe the dependence of a combination of regulatory sites in different functional categories. As shown in Table 9, we select five functional categories as an example; the five MIPS functional categories are “Phosphate Metabolism”, “Glycolysis and Gluconeogenesis”, “Cytokinesis”, “tRNA-Synthetases”, and “Transport ATPases”, to show the differential combinations (R-value exceeding 2.0, at least one support value greater than 0.5) of regulatory sites. The numbers of genes in each functional category are shown in the third row, along with the number of MIPS functional categories. For example, in the first row the combination “tataca, ttgaaa” occurs in the gene upstreams of functional categories, 0104, 0201, 0316, 0510, and 0722, with Chi-square values 4.84, 2.84, 0.30, 0.01, and 1.17, respectively. Chi-square values greater than 3.84 are shown in parentheses. Similarly, the support values are 0.52, 0.00, 0.11, 0.28, and 0.14, respectively. The support values greater than 0.5 are also shown with parentheses. The R-value is shown in the last column and the R-value of the combination “tataca, ttgaaa” is 4.52. Three combina tions of “tataca, ttgaaa”, “aattta, tatata”, and “aattta, tataca” are differential in functional category of “Phosphate Metabolism/0104” then other functional categories by

Table 9. The differential combinations of regulatory sites in five different functional categories (R > 2.0 and support > 0.5).

		Functional Categories (Numeric in MIPS)										
	Number of Genes	0104[a]	0201[b]	0316[c]	0510[d]	0722[e]	0104	0201	0316	0510	0722	
		31	34	90	36	43	31	34	90	36	43	
Category	Differential Combinations	χ^2values					Support values					R values
0104	tataca, ttgaaa	(4.84)	2.84	0.30	0.01	1.17	(0.52)	0.00	0.11	0.28	0.14	4.52
	aattta, tatata	(4.51)	0.05	1.56	0.47	1.71	(0.58)	0.24	0.16	0.14	0.26	3.83
	aattta, tataca	(4.84)	0.01	0.00	1.68	0.47	(0.52)	0.00	0.01	0.08	0.07	5.28
0201	acatat, tatata	0.42	(4.19)	3.56	(8.28)	(5.14)	0.35	(0.53)	0.19	0.14	0.21	2.20
	aatgga, tatata	0.86	(5.28)	0.07	0.29	(4.77)	0.16	(0.56)	0.11	0.11	0.14	6.74
	aatgga, atatat	1.15	(5.72)	0.07	0.04	0.37	0.19	(0.53)	0.04	0.06	0.05	6.94
	TTATC, atggaa	2.40	(5.99)	0.56	0.13	0.81	0.42	(0.56)	0.16	0.17	0.14	2.19
	TATAAA, ttgaaa	0.26	(6.05)	0.26	0.17	0.19	0.39	(0.53)	0.26	0.36	0.12	2.45
	GGGG, TCTCC	0.52	(5.99)	0.04	(4.36)	0.23	0.45	(0.56)	0.00	0.00	0.00	6.39
	GATAA, atggaa	3.32	(5.99)	0.29	0.13	0.81	0.42	(0.56)	0.16	0.17	0.14	2.19
	CCGA, aggaag	0.20	(4.26)	3.80	0.18	1.45	0.32	(0.65)	0.09	0.03	0.16	8.46
0316	GATAA, atttga	0.04	1.88	(10.08)	0.13	1.01	0.39	0.29	(0.53)	0.28	0.37	2.92
	TTATC, atttga	0.62	1.88	(8.30)	0.13	1.01	0.39	0.29	(0.52)	0.28	0.37	2.82
0510	attcaa, ttgaaa	0.26	1.09	2.34	(4.70)	0.19	0.39	0.18	0.36	(0.58)	0.05	6.12
0722	GATAA, tacata	1.01	0.19	3.10	0.01	(3.92)	0.42	0.41	0.48	0.28	(0.65)	3.20
	TTATC, tacata	0.30	0.19	2.20	0.01	(3.92)	0.45	0.41	0.47	0.28	(0.65)	3.12
	aattat, tattaa	0.28	0.17	0.01	0.50	(4.53)	0.26	0.24	0.29	0.36	(0.51)	2.91

Table 10. The differential combinations of regulatory sites in three different functional categories. (R > 2.0 and support > 0.4).

		Functional Categories (Numeric in MIPS)						
		0722[a]	0725[b]	0728[c]	0722	0725	0728	
	Number of ORFs	43	28	35	43	28	35	
Category	Differential Combinations	χ^2values			Support value			R values
0722	atgaat, tattaa	(4.72)	0.96	2.91	(0.44)	0.14	0.09	5.14
	agttaa, atttga	(8.67)	3.19	0.11	(0.42)	0.14	0.03	4.53
0725	atagta, tcatca	0.18	(5.04)	0.00	0.09	(0.43)	0.03	4.21
	aaattg, acatat	2.87	(7.53)	0.09	0.21	(0.57)	0.09	4.50
	TCTCC, aaattg	(4.06)	(6.32)	1.65	0.30	(0.61)	0.00	4.26
	TATAAA, tacgaa	0.02	(6.22)	1.86	0.12	(0.57)	0.09	5.93
	TATAAA, atacgaa	2.88	(4.04)	2.76	0.09	(0.46)	0.00	6.51
	GGGG, tacgaa	2.38	(7.00)	0.72	0.09	(0.54)	0.14	6.41
	GGGG, atacgaa	(4.86)	(8.09)	1.23	0.02	(0.46)	0.00	12.04
0728	atatga, cggaaa	3.80	0.62	(8.07)	0.05	0.18	(0.43)	7.13
	atatatg, tataca	1.74	(7.60)	(8.67)	0.09	0.25	(0.43)	4.63
	atatat, gtatca	1.32	2.59	(7.20)	0.09	0.14	(0.43)	5.15
	aatgtg, tataca	0.32	(5.24)	(11.51)	0.12	0.25	(0.49)	4.85
	aatgtg, atatat	1.34	0.03	(5.54)	0.07	0.21	(0.46)	6.36
	aatgtg, atatac	0.13	1.62	(5.11)	0.09	0.25	(0.43)	4.63

considering the R-values of the three combinations greater then threshold 2.0. All the support values of these categories are also greater than 0.5, while in other categories the support values are very less than 0.5. Similarly, some combinations are also differential in other MIPS functional categories.

Similarly, Table 10 shows another example with three nearly functional categories of transporting genes, i.e., transport ATPases, ABC Transporters, and drug transporters. The differential combinations (R-value exceeding 2.0, at least one support value greater than 0.4) of regulatory sites are also found to investigate the functional-specific combinations in each group of genes.

5. DISCUSSION

This study finds site associations of known regulatory sites and over-represented oligonucleotides located within the promoter regions of groups of functionally related genes. Each promoter region is mapped to a "transaction", and known regulatory sites and over-represented oligonucleotides are mapped to items of a transaction. The data mining techniques are then applied to mine the associations. The enormous number of associations makes it extremely difficult to identify those which are interesting and useful. Finally, the redundant rules are pruned and putative regulatory elements are obtained from the rest of the associations.

Our proposed approach can mine putative functional-specific regulatory elements of any complete genomes such as yeast in this study. The parameters needed to tailor over-represented repetitive sequences within promoter regions of genes can be specified by users according to their needs. The discovered associations of known and putative regulatory elements can also provide effective information to use in studying the mechanisms of gene transcriptional regulation.

Helden *et al.* has developed a method for deciphering the mechanism underlying the common transcriptional response of a set of genes, i.e. discovering cis-acting regulatory elements from a set of unaligned upstream sequences. This method, called dyad analysis, is based on the observation that many regulatory sites consist of a pair of highly conserved tri-nucleotides, spaced by a non-conserved region of fixed width [15]. The transcription factor binding sites in the dyad form are not investigated in our study. We instead focus on the occurrences of combinations of known site homologs and over-represented oligonucleotides in particular gene groups, e.g., MIPS functional categories. The TF binding sites in dyad forms are very important in the identification of transcriptional regulatory sites and we plan to consider the DNA motif prediction approach and dyad analysis for the identification of regulatory sites in the future.

Note that the occurrences of repetitive sequences and known TF binding sites indicate the repetitive elements are putative regulatory elements because groups of transcription factors usually occur cooperatively. By considering the functional-specific occurrence associations of known site homologs and repetitive sequences, the repetitive sequences can be viewed as putative functional-specific 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 in need of being verified by biologists.

ACKNOWLEDGMENTS

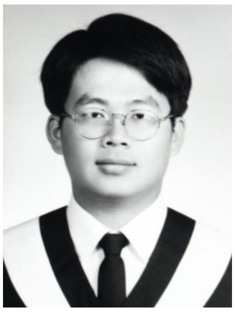
The authors would like to thank the National Science Council of the R.O.C. for financial supporting for this research under Contract No. NSC 89-2213-E-008-061, NSC 90-2213-E-008-053, and NSC 90-2213-E-008-052. In addition, we would like to thank Professors Cheng-Yen Kao at National Taiwan University and Chi-Gong Tong at National Central University for their valuable suggestions and comments.

REFERENCES

1. J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, Vol. 278, 1997, pp. 680-686.
2. 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.
3. A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen, "Predicting gene regulatory elements in silico on a genomic scale," *Genome Res*, Vol. 8, 1998, pp. 1202-1215.
4. 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.
5. 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*, 1997, pp. 65-74.
6. 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," *Silico Biology*, Vol. 2, 2002, pp. 263-273.
7. 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.
8. O. V. Kel-Margoulis, A. G. Romashchenko, N. A. Kolchanov, E. Wingender, and A. E. Kel, "COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation," *Nucleic Acids Res*, Vol. 28, 2000, pp. 311-315.
9. R. Srikant, Q. Vu, and R. Agrawal, "Mining generalized association rules," in *Proceedings of 21st International Conference on Very Large Databases*, 1995, pp. 407-419.
10. 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.
11. H. W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, and B. Weil, "MIPS: a database for genomes and protein sequences," *Nucleic Acids Res*, Vol. 30, 2002, pp. 31-34.
12. D. Gusfield, *Algorithm on Strings*, *Trees*, *and Sequences*, Cambridge University Press, New York, 1997.
13. B. Magasanik, "Regulation of nitrogen utilization," in E. W. Jone, J. R. Pringle, and

J. R. Broach, (eds.), *The Molecular and Cellular Biology of Yeast Saccharomyces: Gene Express*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, Vol. 2, 1992, pp. 283-318.

14. Y. Oshima, N. Ogawa, and S. Harashima, "Regulation of phosphatase synthesis in Saccharomyces cerevisiae – a review," *Gene*, Vol. 179, 1996, pp. 171-177.
15. J. van Helden, A. F. Rios, and J. Collado-Vides, "Discovering regulatory elements in non-coding sequences by analysis of spaced dyads," *Nucleic Acids Res*, Vol. 28, 2000, pp. 1808-1818.



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 from 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. In 2004, he joined the Department of Biological Science and Technology, National Chiao Tung University. His current research interests are Bioinformatics, database systems, and data mining.

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, Chungli, Taiwan, where he became Professor in 2002. His current research interests include database systems, data mining, genetic algorithms, and bioinformatics.

Chia-Hui Chang (張嘉惠) is an Assistant Professor at the Department of Computer Science and Information Engineering, National Central University in Taiwan. She received her B.S. in Computer Science and Information Engineering from National Taiwan University, Taiwan in 1993 and got her Ph.D. in the same department in Jan. 1999. She worked as a post-doctoral in the Institute of Information Science, Academia Sinica after graduation, and then joined National Central University from Aug. 1999. Her research interests include information retrieval and extraction from WWW, pattern mining and knowledge discovery from databases, etc. She has one U.S. patent pending.

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, U.S.A., 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.

Jing-Yue Hong (洪晶月) was born in Nantou, Taiwan, on June 21, 1978. She received M.S. degree in Computer Science and Engineering from Yuan Ze University, Taiwan in 2002. She is interested in the research of bioinformatics, data mining, database systems and communication network.

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 Chungli, 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.