Biological Data Warehousing System for Identifying Transcriptional Regulatory Sites From Gene Expressions of Microarray Data

Ann-Ping Tsou, Yi-Ming Sun, Chia-Lin Liu, Hsien-Da Huang, Jorng-Tzong Horng, Meng-Feng Tsai, and Baw-Juine Liu

***Abstract*—Identification of transcriptional regulatory sites plays an important role in the investigation of gene regulation. For this propose, we designed and implemented a data warehouse to integrate multiple heterogeneous biological data sources with data types such as text-file, XML, image, MySQL database model, and Oracle database model. The utility of the biological data warehouse in predicting transcriptional regulatory sites of coregulated genes was explored using a synexpression group derived from a microarray study. Both of the binding sites of known transcription factors and predicted over-represented (OR) oligonucleotides were demonstrated for the gene group. The potential biological roles of both known nucleotides and one OR nucleotide were demonstrated using bioassays. Therefore, the results from the wet-lab experiments reinforce the power and utility of the data warehouse as an approach to the genome-wide search for important transcription regulatory elements that are the key to many complex biological systems.**

***Index Terms*—Databases, data warehouse, gene expression, gene regulation, microarray, regulatory sites, synexpression group, transcription factor.**

I. Introduction

GENE regulation is one of the most challenging and exciting areas in molecular genetics. Genome-wide gene-expression data provide a unique set of genes and are used to decipher the mechanisms that underlie the common regulations of transcriptional response. The large amount of information gained from the projects for sequencing and elucidating gene expression of the human genome enables researchers to use a computational approach to investigate the mechanism by which genes are regulated.

A transcription factor (TF), which is a DNA-binding protein, can regulate gene expressions and bind to specific sites in the upstream regions of the gene. A variety of TFs, which recognize the specific sites, cooperatively regulate gene transcription by interacting with RNA polymerase. Gene transcription mechanisms can be deciphered by firstly detecting gene regulatory sequences recognized by TFs that regulate the activation of the genes.

Oligo-analysis has been developed to detect over-represented (OR) oligonucleotides in upstream regions. It is based on a systematic counting of occurrences of all possible oligonucleotides in a given sequence [1], [2]. The experimentally identified TF-binding sites were obtained from TRANSFAC (professional 8.3), which contains 14 406 sites and 5711 factors [3].

Three popular regulatory site prediction programs were integrated into the system to discover DNA motifs and, thus, to identify the binding sites in a group of upstream regions. The Gibbs sampler was used [4] with the option "site sampler." One hundred "seeds" or starting points were used; a maximum of 2000 iterations were performed for each run, and the highest scores were reported. The MEME algorithm uses an expectation maximization algorithm for finding patterns in input sequences. AlignACE [6] is based on a Gibbs sampling algorithm and returns a series of motifs that are OR in the upstream regions of the genes of interest.

A previous study of regulatory site prediction by Horng *et al.* presented a data-mining method to detect the associations between site occurrences with combinations of known TF-binding site homologs and OR oligonucleotides [7], [8]. Here, the method is extended to three categories of potentially regulatory sequences. Accordingly, the implemented algorithm detects sites that occur concurrently in the upstream regions of a specified gene group, and also finds the site co-occurrences that have both a support and a confidence value.

RSA-tools [2] is a website for performing computational analysis of regulatory sequences. A suite of computer programs have been developed for the analysis of transcriptional regulatory sequences. The TOUCAN system is a Java application for predicting *cis*-regulatory elements from a set of coexpressed or coregulated genes [9]. TOUCAN does not provide detection of the co-occurrence of regulatory sites.

In this paper, an integrated biological data warehousing system for analyzing transcriptional regulatory sites in the human and mouse genomes was designed and implemented. Users can input a set of gene-expression levels from microarray data, a gene group, or a set of upstream sequences, and then work on the analysis of their transcriptional regulatory sequences in a

Manuscript received October 30, 2003; revised August 18, 2004, February 3, 2005, and April 3, 2005. The work of A.-P. Tsou was supported by Promoting Academic Excellence of Universities Phase II under Contract NSC 96-2752-B-010-002-PAE. The work of H.-D. Huang and J.-T. Horng was supported by the National Science Council, Taiwan, R.O.C., under Contract NSC 93-3112-B-008-002.

A.-P. Tsou and C.-L. Liu are with the Institute of Biotechnology in Medicine, National Yang-Ming University, Taipei 112, Taiwan, R.O.C. (e-mail: aptsou@ym.edu.tw).

Y.-M. Sun, J.-T. Horng, and M.-F. Tsai are with the Institute of Computer Science and Information Engineering, National Central University, Chung-Li 320, Taiwan, R.O.C. (e-mail: felix@db.csie.ncu.edu.tw; horng@db.csie.ncu.edu.tw).

H.-D. Huang is with the Department of Biological Science and Technology, Institute of Bioinformatics, National Chiao-Tung University, Hsinchu 300, Taiwan, R.O.C. (e-mail: bryan@mail.nctu.edu.tw).

B.-J. Liu is with Yuan-Ze University, Chung-Li 320, Taiwan, R.O.C.

Digital Object Identifier 10.1109/TITB.2005.862466

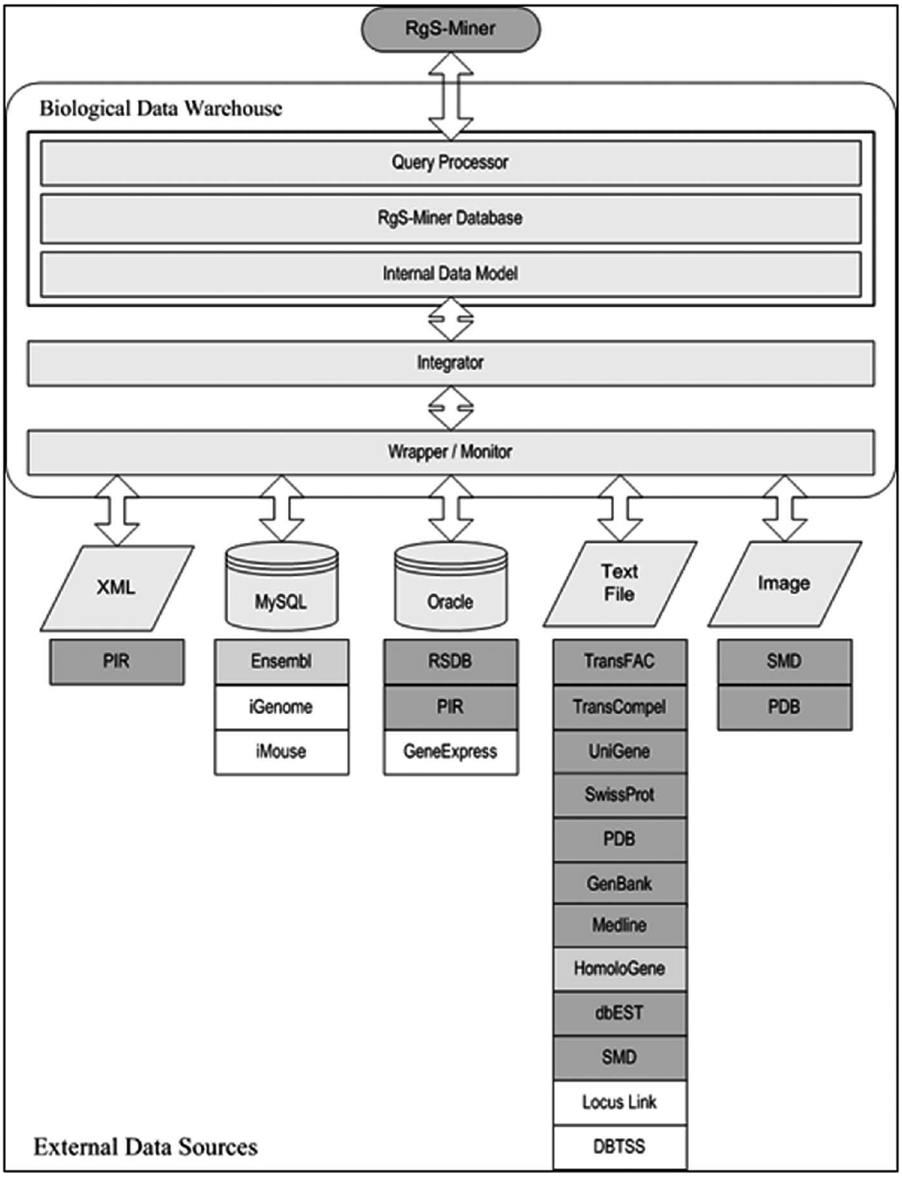


Fig. 1. Biological data warehouse overview. The data warehouse integrates multiple heterogeneous biological data sources from data types such as text-file, XML, image, MySQL database model and Oracle database model. The relational database model is incorporated in the internal database model of the biological data warehouse. Wrappers and monitors are designed for each type of biological database. The wrappers convert the external data into the internal data model. The monitors assess the states of the external data sources and update the internal data. All the external data sources shown in white boxes are newly integrated in this version of RgS-Miner, whereas the data sources in gray boxes were integrated in previous versions.

stepwise manner. The system returns putative regulatory sites, as well as co-occurrences of sites. The specific aim is to develop a predictive system that automatically performs the gene upstream analysis allowing prediction of transcriptional regulatory sites. The predictive system facilitates the detection of regulatory sites in upstream regions of the genes and makes it possible to discover co-occurrence of the regulatory sites. The goal in this work is mainly to establish a biological data warehouse for the computational analysis of transcriptional regulatory sequences in gene upstream sequences. The system facilitates a comprehensive *in silico* gene regulation analysis process for correlating coregulated gene groups from gene-expression profiles, predicting regulatory sites in coregulated gene upstream regions, and detecting the co-occurrence of putative sites.

II. System and Its Implementation

A. System

Since the analysis using this system requires multiple biological data sources, we designed and implemented a data warehouse based on a relational database management system (RDBMS) to integrate and to maintain a variety of heterogeneous biological databases, such as GenBank [10], Ensembl [11], TRANSFAC [12], and so on. The biological data warehouse enables the uniform query interface to access the databases and provides more efficient data management. The data warehousing system is shown in Fig. 1.

We designed and implemented the data warehouse to integrate multiple heterogeneous biological data sources. The relational database model is incorporated in the internal database model of the biological data warehouse. Wrappers and monitors were designed for each type of biological database. The wrappers convert the external data into the internal data model. The monitors assess the states of the external data sources and update the internal data.

To integrate the external data sources into the internal database in the warehousing system, the integrator is responsible for bringing source data into the data warehouse, propagating changes in the source relative to the data warehouse, and maintaining the data extracted in the data warehouse. The wrapper and monitor for each database were designed and implemented. The major tasks of the wrapper and monitor are translation and change detection. The wrapper is responsible for translating the schema of the information source it is concerned with into the schema that is used by the data warehousing system. The monitor module is in charge of detecting any change in the information source it connects to and reporting those changes to the component above, the integrator. Any changes in the information sources will be propagated to the integrator.

RgS-Miner, as described in [13], is a system to analyze transcriptional regulatory sequences. Users first input a set of genes or a set of upstream sequences. The preprocessing phase returns a set of upstream regions. In the subsequent prediction phase, statistical and computational methods, known site matching, detection of OR oligonucleotides, and DNA motif discovery, are provided to predict regulatory sites.

The system then groups the redundant motifs and selects a representative motif for each such group. The annotation phase involves identifying the co-occurrence of regulatory sites following the detection of the putative regulatory sites and motif groups in the prediction phase. For each site found in a particular group of gene upstream regions, a statistical measure, based on the cumulative hypergeometric distribution, is determined to filter out insignificant sites. The putative regulatory sites and site co-occurrences are presented in both textual and graphical formats. The system also considers the evolutionary analyses of transcriptional regulatory sites by using comparative genomics data.

The data warehousing system proposed here also provides a uniform query interface for the easy retrieval of the biological information required in the analysis of the transcriptional regulatory sites in the system. The system enables the following functions: 1) extraction of gene information and tailoring the upstream regions; 2) predicting regulatory sites; 3) detecting site co-occurrences; 4) tools for the visualization of the synergy between TFs; and 5) user profiles and history pages. Additionally, our system integrates multiple regulatory site prediction methodologies and implements an approach to refine the resulting regulatory site into nonredundant ones. The system makes the complicated analyzing processes easier and provides a more user-friendly interface on the web.

TABLE I
DATABASE LINKS IN THE DATA WAREHOUSING SYSTEM

Categories	Database sources	Data type	Ref.
Nucleic acid sequences	GenBank	Text-file	[10]
	Ensembl	MySQL database model	[11]
Genes	GenBank	Text-file	[10]
	Ensembl	MySQL database model	[11]
	SWISS_PROT	Text-file	[14]
	PIR	XML, And Oracle database model	[15]
Gene Expression Profile	UniGene, dbEST,	Text-file	[16], [17]
Repetitive Sequences	RSDB	Oracle database model	[8]
Transcription Factor and Binding Sites	TRANSFAC, TRANSCompel	Text-file	[3], [18]
CpG Islands	HGB	Text-file	[19]
Promoters	Ensembl and Eponine	MySQL database model	[11], [20]
Literature	Medline	Text-file	[21]
Gene Homology	HomoloGene	Text-file	[21]
Microarray Gene expression profiles	The Stanford Microarray Database (SMD)	Text-file, and Images	[22]
Transcriptional Start site	DBTSS	Text-file	[23]

B. System Implementation

The biological data warehouse is implemented by using the MySQL RDBMS version 4.01, which runs on a PC server under the Linux Red Hat 9.0 operating system. The wrapper and monitors are written in the C/C++ programming language. Motivated by the observation that enormous computations are inevitable when working on oligonucleotide analysis to identify regulatory sites in the upstream regions of *Saccharomyces cerevisiae* [1], [2], a more efficient strategy becomes necessary when dealing with the larger eukaryotic genome. Here, we construct the human and mouse genomic sequences into a special computational data structure to reduce the algorithmic complexity when searching for an oligonucleotide in the genomic sequences. Accordingly, the occurrences of a query oligonucleotide are returned efficiently by querying the suffix-array of the considered genome.

We construct the suffix-array to support an efficient way of querying for the occurrences of oligonucleotides whose lengths range from 4 to 25 bps. In our previous study [25], the constructed suffix-array of eukaryotic genome sequence was named i-Genome.

C. External Data Sources

The external data sources required by the system are listed in Table I. Each data source is categorized by its biological meaning and formats. The data types of the external data sources are text-file, XML, image, MySQL database model, and Oracle database model. Generally, most of the external data sources provide data files that can be downloaded freely and directly. The criteria for selecting the specific data sources depend on the materials used in the system.

D. Internal Data Model and Query Processor

The data warehouse can convert the various data formats into the relational database model and store the data into

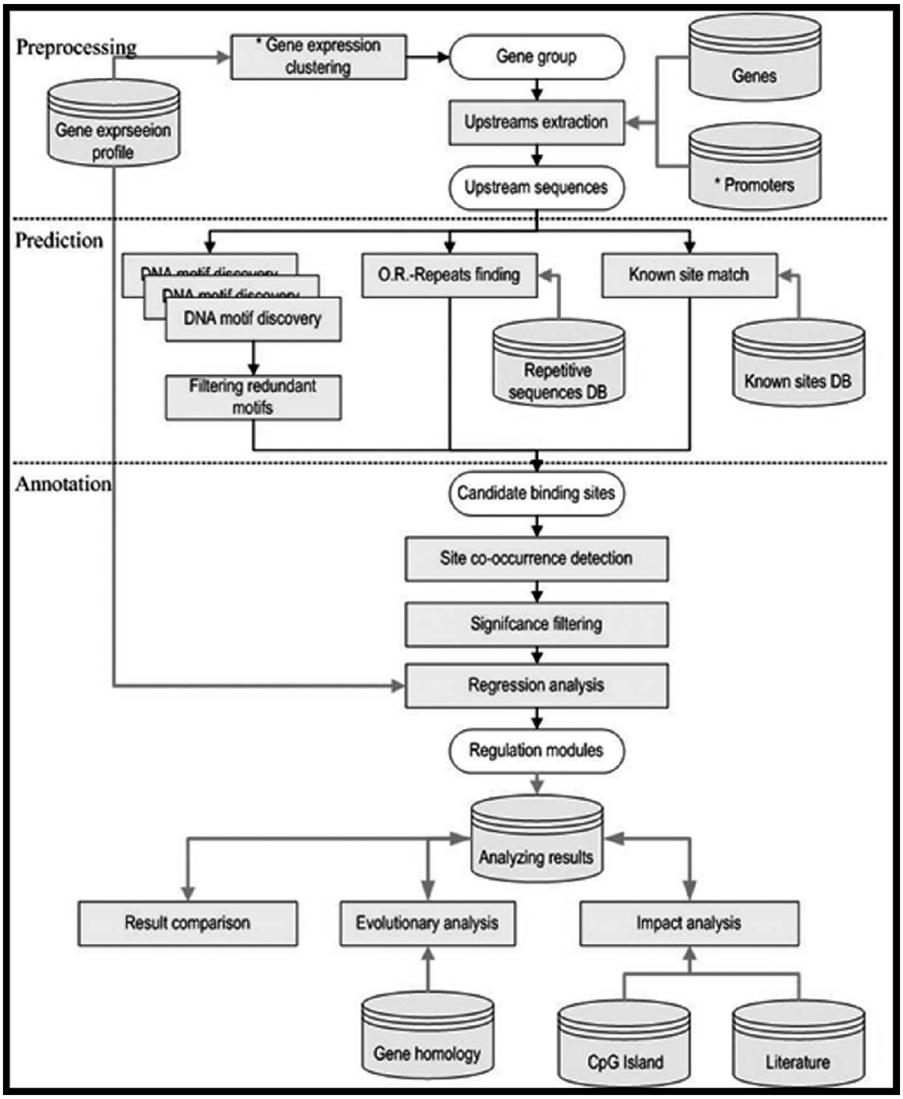


Fig. 2. System flow for analyzing transcriptional regulatory sequences. Users first input a set of genes or a set of upstream sequences. In the prediction phase, statistical and computational methods, known site matching, detection of OR oligonucleotides, and DNA motif discovery, are provided to predict regulatory sites. The annotation phase for identifying the cooccurrence of regulatory sites follows the detection of the putative regulatory sites and motif groups in the prediction phase. The results of the analysis can be annotated and the results from different datasets can be compared. The homologous gene databases can be used for the consideration of evolutionary TF-binding sites. RgS-Miner also links the results of the analysis to literature databases for impact analysis.

the warehouse. The internal database schema is designed to maintain the required biological information from different databases. To maintain the user profiles and by analyzing histories, the RgS-Miner system stores the user input cases and the results from each step of the analysis in the biological data warehouse. The reader may refer to [26] and [27] to find out how the external data sources are integrated under the wrapping rules.

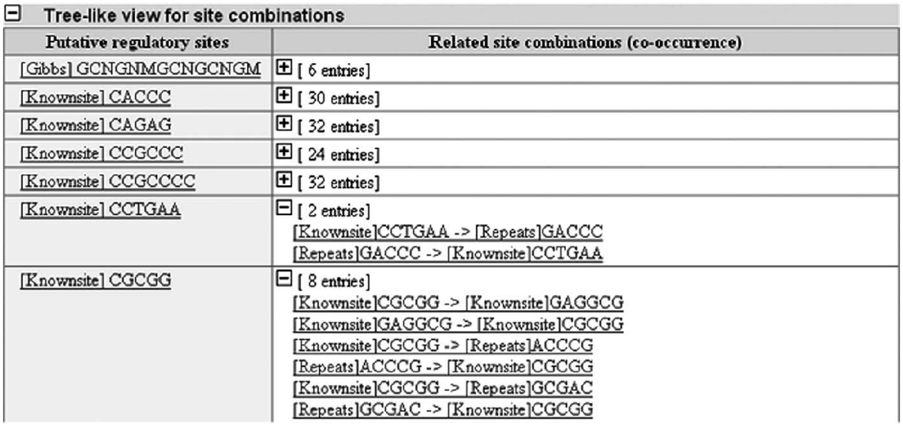
E. Integrator and the Wrapper/Monitor

To integrate the external data sources into the internal database in the warehousing system, the integrator is responsible for bringing source data into the data warehouse, propagating changes in the source relations to the data warehouse, and maintaining the data extracted into the data warehouse. The major tasks of the wrapper/monitor are translation and change detection.

F. System Flow for Identifying Transcriptional Regulatory Sites

Fig. 2 shows the system flow for analyzing transcriptional regulatory sequences. In the preprocessing phase, the gene upstream sequences can be obtained from our database through a query or from user-submitted sequences if the gene instances are not found in the database. In the prediction phase, oligonucleotide analysis, known site matching, and DNA motif discovery tools are applied [13]. The experimentally identified TF-binding sites were obtained from TRANSFAC (professional 8.3).

Three popular regulatory sites prediction programs—a Gibbs sampler, MEME, and AlignACE—were integrated to discover DNA motifs and thus identify the binding sites in a group of upstream regions. The CompareACE score [6], based on the Pearson correlation coefficient between the nucleotide base frequencies of two motif alignments, is used to measure the



⊟ Tree-like view for site combinations

Putative regulatory sites	Related site combinations (co-occurrence)
[Gibbs] GCNGNMGCNGCNGM	⊞ [6 entries]
[Knownsite] CACCC	⊞ [30 entries]
[Knownsite] CAGAG	⊞ [32 entries]
[Knownsite] CCGCCC	⊞ [24 entries]
[Knownsite] CCGCCCC	⊞ [32 entries]
[Knownsite] CCTGAA	⊟ [2 entries] [Knownsite]CCTGAA -> [Repeats]GACCC [Repeats]GACCC -> [Knownsite]CCTGAA
[Knownsite] CGCGG	⊟ [8 entries] [Knownsite]CGCGG -> [Knownsite]GAGGCG [Knownsite]GAGGCG -> [Knownsite]CGCGG [Knownsite]CGCGG -> [Repeats]ACCCG [Repeats]ACCCG -> [Knownsite]CGCGG [Knownsite]CGCGG -> [Repeats]GCGAC [Repeats]GCGAC -> [Knownsite]CGCGG

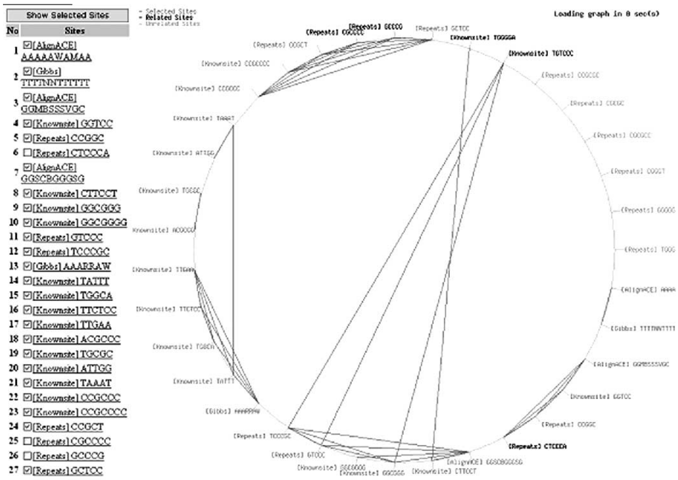
Fig. 3. Web interfaces. A tree-like view to show site combinations.

similarity between pairs of motifs. The occurrence sequences of a motif are used to compute the CompareACE scores. The similarities between each pair of motifs are then used to perform clustering. The K-means clustering method is used to combine similar motifs into groups. The motif groups are used to detect the co-occurrences of sites. The motif group nearest to the centroid of the motif cluster is selected as the representative motif of the motif group.

In the annotation phase, a previous study of regulatory site prediction by Horng *et al.* presented a data-mining method to mine the associations between site occurrences with combinations of known TF-binding site homologs and OR oligonucleotides [7], [8]. That method is herein extended to three categories of potentially regulatory sequences. Accordingly, the implemented algorithm detects sites that occur concurrently in the upstream regions of the gene group of interest, and give the site co-occurrences (also called site combinations) that are found a support value and a confidence value. The cumulative hypergeometric probability distribution has been used to assess the functional significance of computationally derived motifs [6], [28], [29]. In particular, the analyzing results can be further annotated. The results of the analysis of different datasets in different cases can be compared to find the most significant specific regulatory sites in each dataset. The homologous gene databases can be used for the consideration of evolutionary TF-binding sites. RgS-Miner also links the results of the analysis to literature databases for impact analysis.

G. Interfaces

The system can present the results of the analysis in various output formats. It also detects the co-occurrence of putative regulatory sites including known site homologs, OR oligonucleotides, and DNA motif groups as shown in Fig. 3. The output page displays significant site combinations with chi-square values, p-values, support values, confidence values, and the number of occurrences in the relevant upstream regions. The chi-square values and p-values (cumulative hypergeometric probability) are two statistical measurements of the dependencies of the occurrences of the sites in the left part and those of the sites in the right parts of a combination. The positions of occurrences of combinations are depicted graphically on the output pages.



No	Sites
1	☑[AlignACE] AAAAAWAMAA
2	☑[Gibbs] TTTTNNTTTTTT
3	☑[AlignACE] GGMBSSSVGC
4	☑[Knownsite] GGTCC
5	☑[Repeats] CCGGC
6	□[Repeats] CTCCCA
7	☑[AlignACE] GGSCBGGGSG
8	☑[Knownsite] CTTCCT
9	☑[Knownsite] GGCGGG
10	☑[Knownsite] GGCGGGG
11	☑[Repeats] GTCCC
12	☑[Repeats] TCCCGC
13	☑[Gibbs] AAARRAW
14	☑[Knownsite] TATTT
15	☑[Knownsite] TGGCA
16	☑[Knownsite] TTCTCC
17	☑[Knownsite] TTGAA
18	☑[Knownsite] ACGCCC
19	☑[Knownsite] TGCGC
20	☑[Knownsite] ATTGG
21	☑[Knownsite] TAAAT
22	☑[Knownsite] CCGCCC
23	☑[Knownsite] CCGCCCC
24	☑[Repeats] CCGCT
25	□[Repeats] CGCCCC
26	□[Repeats] GCCCG
27	☑[Repeats] GCTCC

Fig. 4. Web interfaces. A circular synergy map of site combinations.

As shown in Fig. 4, the circular synergy map shows the synergism between putative regulatory sites. The circular synergy map is a dynamic web page.

Fig. 5 gives a map that shows the locations of site combinations.

III. Application of the Biological Data Warehouse to Study Transcription Regulation of Synexpression Groups Derived From a Gene-Expression Microarray Analysis

The biological data warehouse is capable of predicting transcriptional regulatory sites of a set of genes that are potentially coregulated. Therefore, it is an ideal tool to study synexpression groups involved in the complex eukaryotic biological system. The development of gene-expression microarray technology for monitoring the transcriptome allows construction of coexpressed and potentially coregulated gene groups. Hence, the gene-expression microarray has been widely used for exploring and integrating the complex processes in normal physiology and in pathology [31]–[34]. We tested the application of the

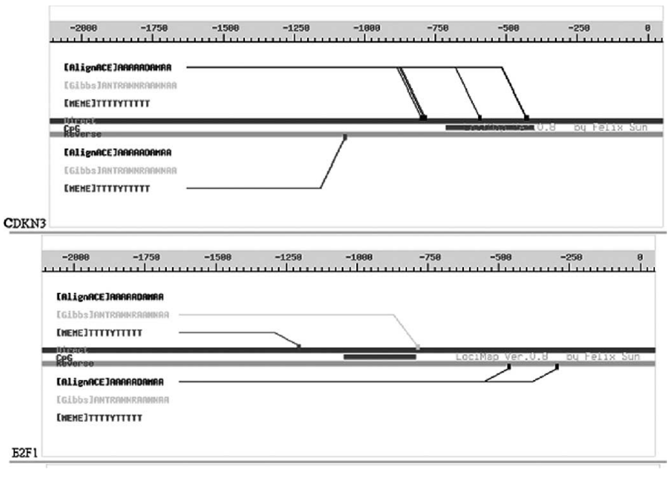


Fig. 5. Web interfaces. The positions of occurrences of combinations are depicted graphically on the output pages.

data warehousing system to identify common transcription regulatory sites in synexpression groups during mouse liver regeneration. The liver is unique among the mammalian organs in its ability to regenerate after severe injury and in disease. Liver regeneration involves three main phases: priming, cell-cycle progression, and tissue remodeling/termination [35]–[37]. Recent microarray analyses [38], [39] have provided new insights into the growth regulation of the liver, but a comprehensive knowledge of the transcription regulation is still lacking.

In this study, we combined *in silico* analysis using RgS-Miner in the data warehouse with wet-lab work to identify the transcriptional regulatory sites in genes coexpressed during the G_2/M phase of the regenerating mouse livers. Previously, we have demonstrated that *Hurp* is expressed during the G_2/M phase of mouse liver regeneration [40] but the transcription regulation of *Hurp* has not been explored. Both the promoter and *cis*-element activities of the *Hurp* gene were investigated.

IV. Experimental Design

A. High-Density Oligonucleotide Microarray Analysis

Liver regeneration was carried out in C57BL/6J mice as described previously [40]. The microarray hybridizations were performed using total RNA from liver samples of control mice (mL) and of mice recovered at various hours after partial hepatectomy (PHx). The GeneChip Mouse Expression 430A array (MOE430A) has 22 690 probesets representing 13 406 UniGenes. Affymetrix MSA 5.0 software was employed to conduct the global scaling normalization, to monitor specific hybridization and gene expression. Both K-means and hierarchical clustering from GeneSpring software version 6.0 (Silicon Genetics, Redwood City, CA) were used for cluster analysis. The relative gene-expression level was defined to be the Log_2 ratio of hybridization intensities between regenerating mouse liver and normal control liver. Genes up- or down-regulated by 1.5-fold at a single or multiple time points were identified. Expression levels of *Ccnb2 Birc5*, *Dlg7* , *Cdca8*, *Tpx2*, *C10orf3*, *Hmmr*, and *mFLJ1090* were verified by RT-PCR. β_2-microglobulin was the internal control.

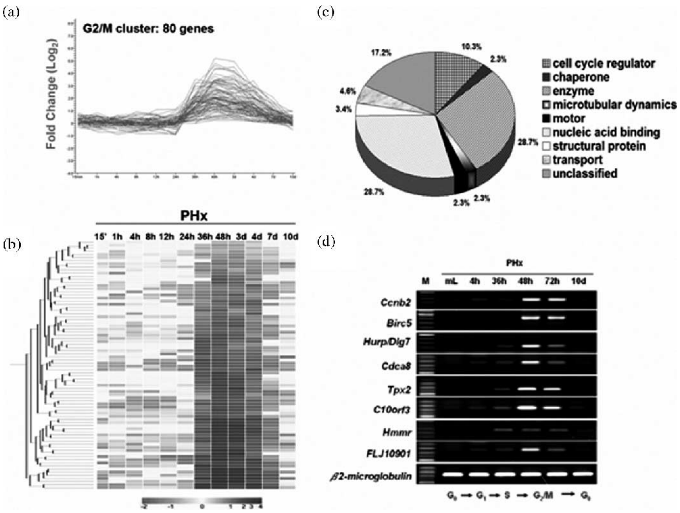


Fig. 6. Coexpressed genes at G_2/M phase of regenerating mouse liver. Cluster analysis of differentially expressed genes identified a group of coexpressed genes during the G_2/M phase (PHx 48–72 h) by (a) K-means clustering and (b) hierarchical grouping using GeneSpring version 6.0. The color scale used to represent the expression ratios (Log2) is shown at the bottom. (c) G_2/M genes were classified by molecular functions. (d) Confirmation of gene expression by RT-PCR. The genes tested are *Ccnb2*, *Birc5*, *Hurp*/*Dlg7*, *Cdca8*, *Tpx2*, *C10orf3*, *Hmmr* and *mFLJ10901*. β_2-Microglobulin is the internal control gene for RT-PCR.

B. Analysis of Promoters and Cis-Elements

Transcription regulation was analyzed using two Luciferase reporters, pGL3-Basic (Promega) and pLuc MCS (Strategene). Genomic DNA fragments surrounding the promoters of *Hurp*, and *Birc5* were subcloned in pGL3-Basic. Two fragments (*Hurp*_I1_R and *Hurp*_I1_L) of the intron 1 of the *Hurp* gene were subcloned in pLuc-MCS for enhancer analysis. Several deletion constructs of the *cis*-element, 5′ cagca 3′, were generated using the QuickChange Site-Directed Mutagenesis Kit (Stratagene). Human 293T cells and HeLa cells were maintained at 37 °C in a 5% CO_2 incubator and grown in DMEM medium supplemented with 10% calf serum and 100 μg/ml penicillin-streptomycin. Calcium phosphate or Metafectene (Biontex) was used in the DNA transfection experiments. HeLa cells were synchronized at G_2/M phase with 50 ng/ml nocodazole treatment (16 h) or at G_1 phase with 400 μM mimosine (16 h). The percentage of cells at different phases of the cell cycle was determined with propidium iodide (400 μg/ml) staining by flow cytometry analysis (FACS analysis, Becton Dickinson FACSort). After 36–48 h, protein lysates were tested for luciferase enzyme activity using a Dual-Luciferase Assay System (Promega).

V. Results and Discussion

A. Coexpressed Genes at G_2/M Phase of Mouse Liver Regeneration

As shown in Fig. 6, a group of 80 genes showed a distinct coexpression pattern at G_2/M phase (PHx 48–72 h) either by the K-means method [Fig. 6(a)] or by hierarchical cluster analysis [Fig. 6(b)]. The majority of the genes belong to cell-cycle regulators (10%), nucleic-acid-binding proteins (28%), or enzymes

TABLE II
PARTIAL KNOWN TF BINDING SITES FOUND BY OUR SYSTEM IN G_2/M PHASE

.	Oligonucleotide	*z*-Score	*p*-Value	Descriptions
1	AAGTGA	64.65	0.000239	HS$IFNB_02/ R00917/ T00422/IRF-1/mouse
2	AGCCAA	73.36	0.000186	MOUSE$NCAM_08/ R01681/ T00537/NF-1/mouse
3	AGGAAA	51.1	0.000383	MOUSE$UPA_01/ R02095/ T00684/PEA3/mouse
4	AGTTCT	59.63	0.000281	MOUSE$RAS1_02/ R01313/ T00335/GR/mouse
5	ATGGGA	63.25	0.000250	MOUSE$AAMY_07/ R01835/ T00701/PTF1-beta/rat
6	ATTGG	48.34	0.000428	MOUSE$M2EAK_08/ R01081/ T00613/NF-Y/mouse
7	CACCC	63.1	0.000251	RAT$TOA_02/ R01474/ T00077/CACCC- factor/human
8	CAGAG	50.8	0.000388	RAT$POMC_03/ R01813/ T00333/GR/rat
9	CAGCAA	62.13	0.000259	MOUSE$THY1_06/ R03046
10	CATTA	39.36	0.000646	HS$GMCSF_03/ R00603/ T00915/YY1/human

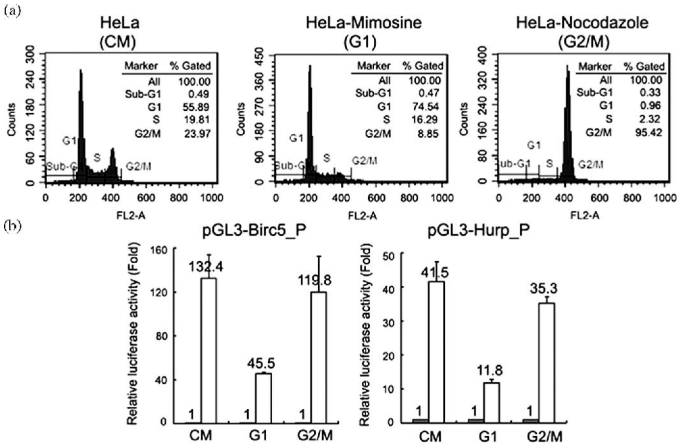


Fig. 7. Hurp promoter is G_2/M regulated. (a) HeLa cells were synchronized with mimosine for G_1 enrichment (75% versus 56% in untreated cells) and with nocodazole for G_2/M enrichment (95% versus 24% in untreated cells). Cell-cycle phase distribution of cells grown in complete medium (CM) served as a control. (b) HeLa cells were transfected with the promoter constructs of Birc5 (pGL3-Birc5_P) or Hurp (pGL3-Hurp_P) before synchronization with mimosine or nocodazole. Relative luciferase activities, expressed as fold difference, were obtained by normalization with the luciferase activity of the vector control (pGL3-Basic), which is marked with a "1." The promoter of both Birc5 and Hurp are highly induced in G_2/M-enriched HeLa cells.

(29%) [Fig. 6(c)] of G_2/M genes, such as *Ccnb2, Birc5* [42], *Hurp*/*Dlg* [40], *Cdca8* [43], *Tpx2* [44] as well as for genes previously unknown to have a G_2/M phase induction such as *C10orf3*, *Hmmr* [45], and *mFLJ10901* [Fig. 6(d)]. RT-PCR analysis demonstrated a good agreement with the microarray analysis.

B. Promoter of Hurp Is Cell-Cycle Regulated

Genomic fragments of *Hurp* were subcloned in pGL3-Basic and promoter activity was assayed in 293T cells. To determine whether these promoters are regulated in a cell-cycle-dependent manner, HeLa cells synchronized at G_1 or G_2/M. In this study, *Birc5* promoter was used as a positive control. As shown in Fig. 7(a), mimosine and nocodazole achieved 75% and 95% synchronization, respectively. The fact that promoters of both *Birc5* and *Hurp* are more active in G_2/M-enriched cells than in G_1-enriched HeLa cells [Fig. 7(b)] suggested that the *Hurp* promoter is cell-cycle regulated.

C. Characterization of the cis-Element of the Hurp Gene

TABLE III
PARTIAL OVER-REPRESENTED REPEAT ELEMENTS DISCOVERED BY OUR SYSTEM IN G_2/M PHASE

	Oligonucleotide	*z*-Score	*p*-Value	Expect	SD
1	aaccg	9.36	0.011418	40.5	6.36
2	agcgg	21.43	0.002178	40.9	6.40
3	caccg	10.92	0.008384	48.2	6.94
4	cagca	3.43	0.084838	493.8	22.20
5	ccaatc	3.35	0.088998	59.2	7.69
6	ccggg	21.52	0.002159	63.1	7.94
7	ccgtc	14.15	0.004994	37.4	6.12
8	ctccg	18.01	0.003084	57.0	7.55
9	ctgcgg	13.69	0.005334	14.6	3.82
10	gaccg	13.18	0.005759	29.5	5.43

We further examined the architecture of *cis*-elements in G_2/M genes. We submitted 3-Kb sequences upstream of the first ATG of the ORFs for regulatory site prediction by the data warehousing system. Data prediction for both known TF-binding sites (Table II) and OR oligonucleotides (Table III) was collected. Partial known TF-binding sites found by our system and partial OR elements discovered by the system are shown in Tables II and III, respectively. Multiple putative TF-binding sites and OR were identified. TF-binding sites common to 80%–95% of the G_2/M genes included Sp1, CREB, CDE/CHR, and NF-Y (data not shown). Multiple copies (five or more) of one specific OR repeat, 5′ cagca 3′, was found in 79 of 80 G_2/M genes. The occurrence of cagca within the 4000 bp upstream of the first ATG of G_2/M genes is significantly higher than in the whole genome (z score $= 3.43, p = 0.0848$) and is also higher than within the 4000-bp upstream regions of genes up-regulated at 96 h after partial hepatectomy (z score $= 0.62$ versus PHx 96 h genes). This 5-bp OR seems to be selective for the genes up-regulated during G_2/M phase. We named this 5-bp nucleotide the Y-like element. The *Hurp* gene has five copies of the Y-like element on both strands of intron 1 [Fig. 8(a)]. To determine the biological roles of these *cis*-elements in the G_2/M genes, an enhancer reporter assay was performed using the *Hurp* gene as the example. Deletion of the core sequence of NF-Y (ccaat) resulted in a reduction of transcription activation by threefold [upper panel, Fig. 8(b)]. NF-Y is a key TF for cell-cycle genes and acts by pre-setting the promoter architecture for other regulatory proteins [46]. Our data also support its critical role as an enhancer for G_2/M genes. Deletion of Y-2 and Y-4 resulted in an almost complete loss of enhancer activity by the *Hurp* gene [Fig. 8(b)]. While the Y-like element (cagca) is a predicted OR, it has to be demonstrated if it binds a specific TF.

Computational methods and the data warehousing system provide a high-throughput means to allow construction of

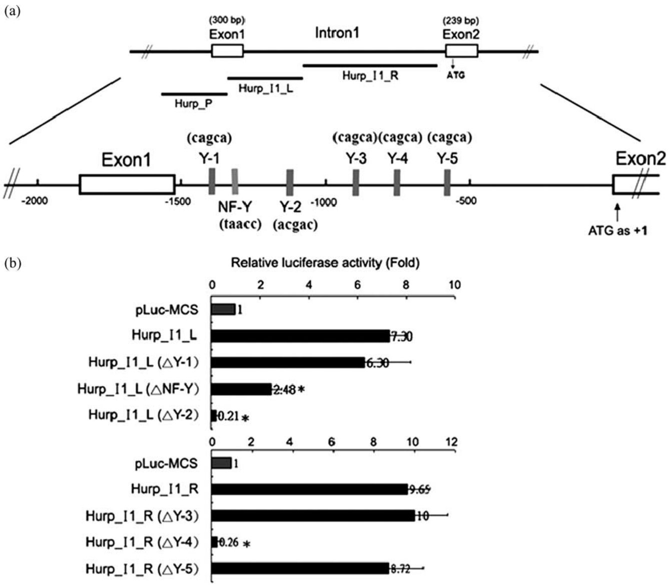


Fig. 8. Detection of *cis*-element activities in intron 1 of *Hurp* gene. (a) Diagram of the *Hurp* gene with the ORF starting in exon 2. Hurp_P harbors the promoter activity as shown in Fig. 3(b). Hurp_I1-L and Hurp_I1-R are fragments encompassing a portion of exon 1 and the entire intron 1. The predicted binding site sequences are marked for intron 1 region: 5' ccaat 3' for NF-Y and 5' cagca 3' for Y-like element. Four Y-like elements (Y-1, Y-3, Y-4, Y-5) are on the sense strand while one Y-like element (Y-2) and the NF-Y site are on the antisense strand. (b) The activity of the *cis*-elements was detected by the Luciferase reporter assay. Deletion of the core sequence of NF-Y, Y-2, or Y-4 resulted in a severe reduction in enhancer activity while deletion of Y-1, Y-3, or Y-5 did not affect the enhancer activity. Relative luciferase activities, expressed as fold difference, were obtained by normalization with the luciferase activity of the vector control (pGL3-Basic) (pLuc-MCS).

regulatory modules for coexpressed genes. Incorporation of comprehensive microarray datasets will further facilitate deciphering the regulatory control mechanisms that govern synexpression groups and their associated molecular pathways critical to complex biological systems.

References

[1] S. Levy, S. Hannenhalli, and C. Workman, "Enrichment of regulatory signals in conserved non-coding genomic sequence," *Bioinformatics*, vol. 17, pp. 871–877, 2001.

[2] J. van Helden, B. Andre, and J. Collado-Vides, "A web site for the computational analysis of yeast regulatory sequences," *Yeast*, vol. 16, pp. 177–187, 2000.

[3] 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," *Nucl. Acids Res.*, vol. 29, pp. 281–283, 2001.

[4] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, "Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment," *Science*, vol. 262, pp. 208–214, 1993.

[5] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 1994, vol. 2, pp. 28–36.

[6] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church, "Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces Cerevisiae*," *J. Mol. Biol.*, vol. 296, pp. 1205–1214, 2000.

[7] 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 Biol.*, vol. 2, pp. 263–273, 2002.

[8] 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," *J. Comput. Biol.*, vol. 9, pp. 621–640, 2002.

[9] S. Aerts, G. Thijs, B. Coessens, M. Staes, Y. Moreau, and B. De Moor, "Toucan: Deciphering the cis-regulatory logic of coregulated genes," *Nucl. Acids Res.*, vol. 31, pp. 1753–1764, 2003.

[10] K. D. Pruitt and D. R. Maglott, "RefSeq and LocusLink: NCBI gene-centered resources," *Nucl. Acids Res.*, vol. 29, pp. 137–140, 2001.

[11] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp, "The Ensembl genome database project," *Nucl. Acids Res.*, vol. 30, pp. 38–41, 2002.

[12] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, "TRANSFAC: Transcriptional regulation, from patterns to profiles," *Nucl. Acids Res.*, vol. 31, pp. 374–378, 2003.

[13] H. D. Huang, J. T. Horng, Y. M. Sun, A. P. Tsou, and S. L. Huang, "Identifying transcriptional regulatory sites in the human genome using an integrated system," *Nucl. Acids Res.*, vol. 32, pp. 1948–1956, 2004.

[14] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000," *Nucl. Acids Res.*, vol. 28, pp. 45–48, 2000.

[15] C. H. Wu, L. S. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R. S. Ledley, B. E. Suzek, C. R. Vinayaka, J. Zhang, and W. C. Barker, "The Protein Information Resource," *Nucl. Acids Res.*, vol. 31, pp. 345–347, 2003.

[16] M. S. Boguski, T. M. Lowe, and C. M. Tolstoshev, "dbEST–database for 'expressed sequence tags'," *Nat. Genet.*, vol. 4, pp. 332–333, 1993.

[17] G. D. Schuler *et al.*, "A gene map of the human genome," *Science*, vol. 274, pp. 540–546, 1996.

[18] O. V. Kel-Margoulis, A. E. Kel, I. Reuter, I. V. Deineko, and E. Wingender, "TRANSCompel: A database on composite regulatory elements in eukaryotic genes," *Nucl. Acids Res.*, vol. 30, pp. 332–334, 2002.

[19] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent, "The UCSC genome browser database," *Nucl. Acids Res.*, vol. 31, pp. 51–54, 2003.

[20] U. Ohler and H. Niemann, "Identification and analysis of eukaryotic promoters: Recent computational approaches," *Trends Genet.*, vol. 17, pp. 56–60, 2001.

[21] D. L. Wheeler, D. M. Church, S. Federhen, A. E. Lash, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, E. Sequeira, T. A. Tatusova, and L. Wagner, "Database resources of the national center for biotechnology," *Nucl. Acids Res.*, vol. 31, pp. 28–33, 2003.

[22] J. Gollub, C. A. Ball, G. Binkley, J. Demeter, D. B. Finkelstein, J. M. Hebert, T. Hernandez-Boussard, H. Jin, M. Kaloper, J. C. Matese, M. Schroeder, P. O. Brown, D. Botstein, and G. Sherlock, "The stanford microarray database: Data access and quality assessment tools," *Nucl. Acids Res.*, vol. 31, pp. 94–96, 2003.

[23] Y. Suzuki, R. Yamashita, S. Sugano, and K. Nakai, "DBTSS, DataBase of Transcriptional Start Sites: Progress report 2004," *Nucl. Acids Res.*, vol. 32, pp. D78–D81, 2004.

[24] D. Gusfield, *Algorithms on Strings, Trees and Sequences*, New York: Cambridge Univ. Press, 1997.

[25] F. M. Lin, H. D. Huang, Y. C. Chang, and J. T. Horng, "i-Genome: A database to summarize oligonucleotide data in genomes," *BMC Genomics*, vol. 5, p. 78, 2004.

[26] J. T. Horng and C. W. Chen, "A mechanism for view consistency in a data warehousing system," *J. Syst. Softw.*, vol. 56, pp. 23–37, 2001.

[27] J. T. Horng and J. Lu, "Modularized design for wrappers/monitors in data warehouse systems," *J. Syst. Softw.*, vol. 3, pp. 185–199, 2000.

[28] L. J. Jensen and S. Knudsen, "Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation," *Bioinformatics*, vol. 16, pp. 326–333, 2000.

[29] P. Sudarsanam, Y. Pilpel, and G. M. Church, "Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*," *Genome Res.*, vol. 12, pp. 1723–1731, 2002.

[30] T. M. Phuong, D. Lee, and K. H. Lee, "Regression trees for regulatory element identification," *Bioinformatics*, vol. 20, pp. 750–757, 2004.

[31] P. A. Clarke, R. te Poele, R. Wooster, and P. Workman, "Gene expression microarray analysis in cancer biology, pharmacology, and drug development: Progress and potential," *Biochem. Pharmacol.*, vol. 62, pp. 1311–1336, 2001.

[32] V. R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. Lee, J. M. Trent, L. M. Staudt, J. Hudson, Jr., M. S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P. O. Brown, "The transcriptional program in the response of human fibroblasts to serum," *Science*, vol. 283, pp. 83–87, 1999.

[33] H. Okabe, S. Satoh, T. Kato, O. Kitahara, R. Yanagawa, Y. Yamaoka, T. Tsunoda, Y. Furukawa, and Y. Nakamura, "Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: Identification of genes involved in viral carcinogenesis and tumor progression," *Cancer Res.*, vol. 61, pp. 2129–2137, 2001.

[34] Z. Xiang, Y. Yang, X. Ma, and W. Ding, "Microarray expression profiling: Analysis and applications," *Curr. Opin. Drug Discov. Devel.*, vol. 6, pp. 384–395, 2003.

[35] N. Fausto, "Liver regeneration," *J. Hepatol.*, vol. 32, pp. 19–31, 2000.

[36] A. Zimmermann, "Liver regeneration: The emergence of new pathways," *Med. Sci. Monit.*, vol. 8, pp. RA53–RA63, 2002.

[37] D. Mangnall, N. C. Bird, and A. W. Majeed, "The molecular physiology of liver regeneration following partial hepatectomy," *Liver Int.*, vol. 23, pp. 124–138, 2003.

[38] M. Arai, O. Yokosuka, T. Chiba, F. Imazeki, M. Kato, J. Hashida, Y. Ueda, S. Sugano, K. Hashimoto, H. Saisho, M. Takiguchi, and N. Seki, "Gene expression profiling reveals the mechanism and pathophysiology of mouse liver regeneration," *J. Biol. Chem.*, vol. 278, pp. 29813–29818, 2003.

[39] Y. Fukuhara, A. Hirasawa, X. K. Li, M. Kawasaki, M. Fujino, N. Funeshima, S. Katsuma, S. Shiojima, M. Yamada, T. Okuyama, S. Suzuki, and G. Tsujimoto, "Gene expression profile in the regenerating rat liver after partial hepatectomy," *J. Hepatol.*, vol. 38, pp. 784–792, 2003.

[40] A. P. Tsou, C. W. Yang, C. Y. Huang, R. C. Yu, Y. C. Lee, C. W. Chang, B. R. Chen, Y. F. Chung, M. J. Fann, C. W. Chi, J. H. Chiu, and C. K. Chou, "Identification of a novel cell cycle regulated gene, HURP, overexpressed in human hepatocellular carcinoma," *Oncogene*, vol. 22, pp. 298–307, 2003.

[41] G. M. Higgins and R. M. Anderson, "Experimental pathology of the liver. Restoration of the liver of white rat following partial surgical removal," *Arch. Pathol.*, vol. 12, pp. 186–202, 1931.

[42] F. Li and D. C. Altieri, "The cancer antiapoptosis mouse survivin gene: Characterization of locus and transcriptional requirements of basal and cell cycle-dependent expression," *Cancer Res.*, vol. 59, pp. 3143–3151, 1999.

[43] R. Gassmann, A. Carvalho, A. J. Henzing, S. Ruchaud, D. F. Hudson, R. Honda, E. A. Nigg, D. L. Gerloff, and W. C. Earnshaw, "Borealin: A novel chromosomal passenger required for stability of the bipolar mitotic spindle," *J. Cell Biol.*, vol. 166, pp. 179–191, 2004.

[44] T. Wittmann, M. Wilm, E. Karsenti, and I. Vernos, "TPX2, a novel xenopus MAP involved in spindle pole organization," *J. Cell Biol.*, vol. 149, pp. 1405–1418, 2000.

[45] C. Fieber, R. Plug, J. Sleeman, P. Dall, H. Ponta, and M. Hofmann, "Characterisation of the murine gene encoding the intracellular hyaluronan receptor IHABP (RHAMM)," *Gene*, vol. 226, pp. 41–50, 1999.

[46] G. Caretti, V. Salsi, C. Vecchi, C. Imbriano, and R. Mantovani, "Dynamic recruitment of NF-Y and histone acetyltransferases on cell-cycle promoters," *J. Biol. Chem.*, vol. 278, pp. 30435–30440, 2003.

Ann-Ping Tsou was born in Tainan, Taiwan, R.O.C., on January 19, 1950. She received the Ph.D. degree in microbiology from Harvard School of Public Health, Boston, MA, in 1982.

From 1985 to 1993, she was a Staff Scientist in the Discovery Research Division, Syntex (USA) Inc., In 1993, she joined the Institute of Biotechnology in Medicine, Taipei, Taiwan, R.O.C., where she is an Associate Professor. Her current research interests include transcription regulation and functional genomics of hepatocellular carcinoma.

Yi-Ming Sun was born in Taitong, Taiwan, R.O.C., in 1972. He received the M.S. degree in computer science and information engineering from National Central University, Chung-Li, Taiwan, R.O.C. in 2003, where he is currently working toward the Ph.D. degree in computer science and information engineering.

His current research interests are bioinformatics and database systems.

Chia-Lin Liu was born in Kaohsiung, Taiwan, R.O.C., in 1979. He received the M.S. degree from the Institute of Biotechnology in Medicine, National Yang-Ming University, Taipei, Taiwan, R.O.C. in 2004.

He is currently with the Institute of Biotechnology in Medicine, National Yang-Ming University.

Hsien-Da Huang was born in Taoyuan, Taiwan, R.O.C., in 1975. He received the Ph.D. degree in computer science and information engineering from National Central University, Chung-Li, Taiwan, R.O.C. in 2003.

In 2003, he joined the Department of Biological Science and Technology and Institute of Bioinformatics, National Chiao-Tung University, Hsinchu, Taiwan. His current research interests include bioinformatics, database systems, and data mining.

Jorng-Tzong Horng was born in Nantou, Taiwan, R.O.C., on April 10, 1960. He received the Ph.D. degree in computer science and information engineering from the National Taiwan University, Taipei, Taiwan, in April 1993.

In 1993, he joined the Department of Computer Science and Information Engineering, National Central University, Chung-Li, Taiwan, where he became a Professor in 2002. His current research interests include database systems, data mining, genetic algorithms, and bioinformatics.

Meng-Feng Tsai received the Ph.D. degree from the University of California, Los Angeles.

He is an Assistant Professor in the Department of Computer Science and Information Engineering, National Central University, Chung-Li, Taiwan, R.O.C. His current research interests include data mining, data-warehousing, and distributed systems.

Baw-Juine Liu received the Ph.D. degree from National Taiwan University, Taipei, Taiwan, R.O.C.

He is a Professor in the Graduate School of Social Informatics, Yuan Ze University, Chung-Li, Taiwan. His current research interests include database systems, object-oriented programming, and bioinformatics.