

# RgS-Miner: A Biological Data Warehousing, Analyzing and Mining System for Identifying Transcriptional Regulatory Sites in Human Genome

Yi-Ming Sun<sup>1</sup>, Hsien-Da Huang<sup>2</sup>, Jorng-Tzong Horng<sup>1,3</sup>,  
Ann-Ping Tsou<sup>4</sup>, and Shir-Ly Huang<sup>3</sup>

<sup>1</sup> Department of Computer Science and Information Engineering  
National Central University, Chung-Li 320, Taiwan  
{horng, felix}@db.csie.ncu.edu.tw

<sup>2</sup> Department of Biological Science and Technology, Institute of Bioinformatics  
National Chiao Tung University, Hsin-Chu 300, Taiwan  
bryan@mail.NCTU.edu.tw

<sup>3</sup> Department of Life Science, National Central University, Chung-Li 320, Taiwan

<sup>4</sup> Institute of Biotechnology in Medicine, National Yang-Ming University  
Taipei 112, Taiwan

**Abstract.** Recently, biological databases and analytical methods have become available for analyzing gene expression and transcriptional regulatory sequences. However, users must make the complicated analyses to query the data in various databases, and then they must analyze the gene upstreams using various predictive tools, before finally converting data among formats. Beyond methods for predicting transcriptional regulatory sites, new automated and integrated methods for analyzing gene upstream sequences on a higher level are urgently required. Efficient and integrated data management methods are essential, too. We present an integrated system, namely RgS-Miner, to predict transcriptional regulatory sites and detect co-occurrence of these regulatory sites. RgS-Miner comprises a biological data warehousing system, pattern discovery programs, pattern occurrence association detectors and user interfaces. The system is available at <http://rgsminer.csie.ncu.edu.tw/>.

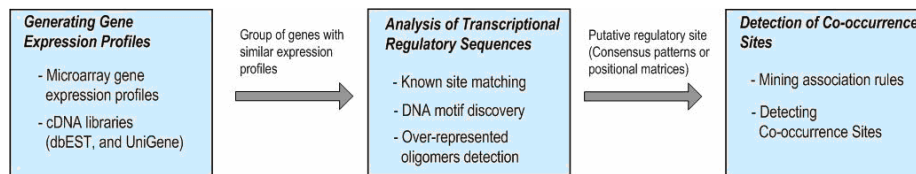
## 1 Introduction

Genome-wide gene expression data provides a unique set of genes and are used to decipher the mechanisms that underlie the common regulations of transcriptional response. Gene regulation is one of the most challenging and exciting areas in molecular genetics. 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 investigating the mechanism by which genes are regulated. Not only do the identification of transcription factor (TF) binding sites yield valuable information on gene expression and regulation, but also detection of the co-occurrence of regulatory sites facilitates the determination of regulation mechanisms. Recently, biological information and analytical methods have become

available for analyzing gene expression and transcriptional regulatory sequences. However, users must make the complicated analyses to query the data in various databases, and then they must analyze the gene upstreams using various predictive tools, before finally converting data among formats. Beyond methods for predicting transcriptional regulatory sites, new automated and integrated methods for analyzing gene upstream sequences on a higher level are urgently required. Identifying regulatory sites requires many biological databases, so efficient and integrated data management methods are essential.

Generally speaking, the analyzing processes for the investigation of the gene transcriptional regulations are mainly described as shown in Fig. 1, the analysis for gene expression profiles are considered to be potentially co-regulated. Intuitively, the analysis of regulatory sequences searches in the upstreams of co-regulated genes for highly conserved patterns which are possible to be regulatory sites. The co-occurrences of putative regulatory sites are detected to decipher the cooperation or synergisms between transcription factors.

An integrated system for analyzing transcriptional regulatory sites in the human genome was designed and implemented. Users can input a gene group or a set of upstream sequences, and then work on the analysis of the transcriptional regulatory sequences stepwise. 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 to predict transcriptional regulatory sites. The predictive system facilitates the detection of regulatory sites in upstream regions of the genes and help to discover co-occurrence of the regulatory sites.



**Fig. 1.** The analyzing process in gene transcriptional regulation

## 2 Designs Goals

The research goal in this work is mainly to establish a predictive system, namely RgS-Miner, for the analyses of transcriptional regulatory sequences in the gene upstream sequences. The system facilitates the comprehensive *in silico* gene regulation analyzing processes of correlating co-regulated gene groups from gene expression profiles, predicting regulatory sites in co-regulated gene upstreams, and detecting the co-occurrence of putative sites.

Since the analyses in the system require multiple biological databases in various types of data sources, a biological data warehouse based on a Relational Database Management System (RDBMS) is designed and constructed to integrate and maintain a variety of heterogeneous biological databases.

The system enables the functions as follows. (1) Extraction of gene information and tailoring the upstream regions. (2) Predicting regulatory sites. (3) Detecting site co-occurrences. (4) The visualization tools showing the synergy between transcription factors. (5) User profiles and historical pages. Additionally, RgS-Miner integrates multiple regulatory site prediction methodologies and proposes an approach to combine the result regulatory site into non-redundant ones.

Another design goal in this work is to design and implement the interface of the RgS-Miner system. Users can input a set of gene groups in the data input page to access the biological data warehouse to obtain the biological information. All the analyzing steps are processed on the web interface step by step and the results of each step are stored in the data warehouse. The system provides both text formats and visualizing formats to show the analyzed results.

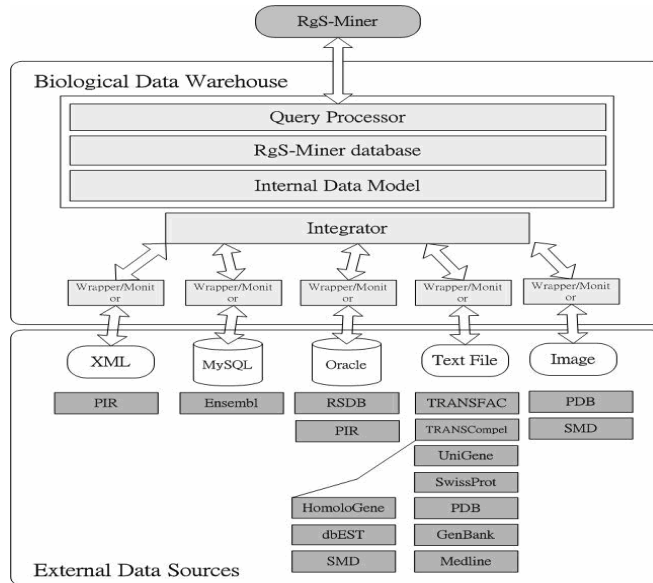
### 3 Architectural Overview of RgS-Miner

The system is designed to record information about human genomic sequences, gene information, the transcription factors (TF) and TF binding sites, gene transcriptional start sites (TSS), repetitive elements, and CpG islands in the database. The system facilitates the detection of regulatory sites in human genome by inputting a gene group, which can be constructed based on the cluster analysis of gene expression data or the genes considered potentially co-regulated under particular transcriptional regulation mechanisms. Additionally, graphical web interfaces are designed to show the upstream regions where regulatory sites or site combinations occurs. User profiles and analyzing histories are also maintained in the database.

#### 3.1 Biological Data Management

A lot of biological databases in different formats provide valuable information in each molecular biology research field. In order to efficiently manage the information from multiple biological databases to facilitate the implementation of the proposed system, we incorporate the concept of data warehousing system to construct a biological data warehouse, which maintains, updates and integrates all the required biological databases in the proposed system. Especially, this study incorporates the repetitive sequences in eukaryotic genomes to detect over-represented repeats during the analysis of transcriptional regulation of gene expression.

As shown in Fig. 2, we design and implement a data warehouse to integrate the RgS-Miner databases and multiple heterogeneous biological data sources such as GenBank at NCBI [1], Ensembl [2], TRANSFAC [3], Eponine [4], and Tandem-Repeat-Finder [5]. 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, which the wrappers are capable of converting the external data into the internal data model and the monitors monitor and update the states of the external data sources. The proposed data warehousing system also provides a uniform query interface for easily retrieving the biological information required in the analysis of transcriptional regulatory sites in RgS-Miner system.



**Fig. 2.** The biological data warehouse overview

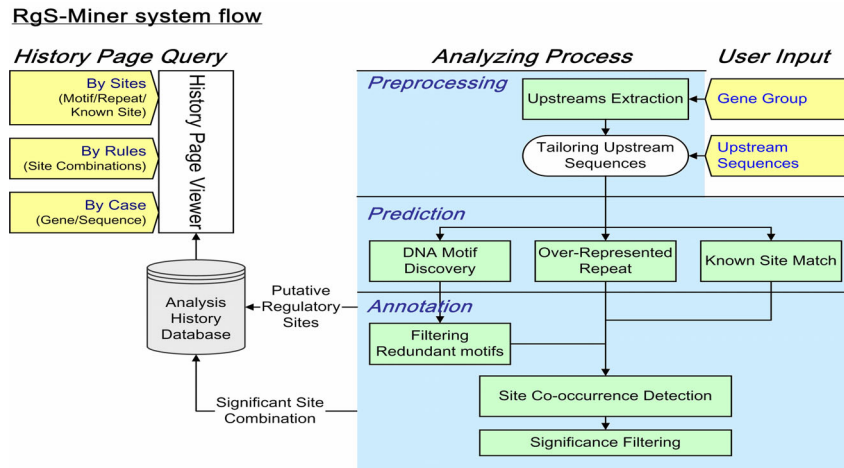
Each external data sources are categorized into different types in biological meaning, as well as the different storing data types. The data types of the external data sources are text-file, XML, image, MySQL database model, and Oracle database model. Especially, some external data sources contain more than one data types, e.g., the protein structures in the protein data bank (PDB) is in text-file as well as structural images. Generally, most of the external data sources provide the data files which can be downloaded freely and directly.

The data warehouse can convert various data formats into the relational database model and store the data into the warehouse. The internal database schema based on relational database model is designed to maintain the required biological information from different databases. To maintain the user profiles and analyzing histories, the RgS-Miner system stores the user input cases and the analyzing results of each steps in the database in the biological data warehouse. For each analyzing case, users can submit a gene group for the analysis of regulatory sequences. The case descriptions, putative regulatory sites, and site co-occurrences are stored in the relations of “Site”, “Cases”, “Patterns”, and “Rules”, respectively.

In order to integrate the external data sources into the internal database in the warehouse 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 in the data warehouse, which may include merging, filtering and summarizing information from different information sources. When storing integrated data, it may need to obtain further information from the same or different information sources. Then it would send requests to the appropriate wrapper modules below it.

The wrapper/monitor for each biological database are designed and implemented. The major tasks of the wrapper/monitor are the translation and change detection. The

wrapper is responsible for translating the schema of the information source it concerns to the schema which is used by the data warehouse system. The monitor module is in charge of detecting any change from the information source it connects to, and reporting those changes to the component above, the integrator. Any change from information sources will be propagated to the integrator.



**Fig. 3.** The system flow

### 3.2 System Flow

Fig. 3 shows the system flow for analyzing 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 over-represented (OR) oligonucleotides and DNA motif discovery, are provided to predict regulatory sites. Users separately run each predictive method to detect the regulatory sites in the upstream regions. Many highly similar regulatory motifs are thus detected. The system has the function to group the redundant motifs; a representative motif is selected in each such group.

The annotation phase for identifying the co-occurrence of regulatory sites follows the detection of the putative regulatory sites and motif groups in the prediction phase. For each site found in a particular group of gene upstreams, a statistical measurement, the cumulative hyper-geometric 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 results of each step of the analysis are automatically stored in the database. User can login the system to query user profiles and history pages, which are then displayed on the web pages.

### 3.3 Preprocessing

The gene upstreams can be from our database through a query or from user submitted sequences if the gene instances are not found in the database. Since some of the genes without the annotation of the predicted TSS, the users can tailor the upstream regions by referring the gene coding region start positions. To prepare the user specified region of gene upstreams, formats described genes including gene symbol, and GenBank accession number are parsed and the upstreams are extracted by querying in the RgS-Miner database. However, the gene upstreams can be from the database or from user submitted sequences if the gene instances are not found in the database. For each gene in the input gene groups, the identifiers passed from previous step are used to retrieve the upstream sequences in DNA alphabet set {A, T, C, G} from the biological data warehousing system.

### 3.4 Regulatory Site Prediction

Olio-analysis has been developed to detect over-represented oligonucleotides in upstream regions. It is based on a systematic counting of occurrences of all possible oligonucleotides in a given sequence [6]. An advantage of the method is that it can detect all the over-represented patterns of a given length in a single run. The system applies a statistical method to discover statistically significant oligonucleotides, which are DNA sequences of small length within the upstream regions of genes by comparing their frequencies of occurrence to their background frequencies of occurrence throughout human genome: the frequencies of occurrence of oligonucleotides yield a pre-constructed index.

The experimentally identified transcription factor binding sites were obtained from TRANSFAC (professional 5.4), which contains 11,537 sites and 4,774 factors [3]. In the system, 3,294 vertebral TF binding sites are matched to upstreams of human genes. A program is implemented to match the consensus patterns of the TRANSFAC known sites to the upstream sequences. The program allows mismatching by considering a mismatch penalty. The known TF binding sites are matched to the prepared upstreams both in double strands; the positions of each known site homologues are stored in the database for further analysis in the annotation phase.

Three popular regulatory sites prediction program - Gibbs sampler [7], MEME [8] and AlignACE [9] – were integrated to discover DNA motifs and thus identify the binding sites in a group of upstream regions. The motifs obtained by the DNA motif discovery methods are stored in the format of consensus pattern, which includes the site sequences that occur in each upstream region.

### 3.5 Filtering out Redundant Regulatory Motifs

Some of the DNA motifs detected by various approaches are highly similar to each other and so are redundant for further analysis for detecting site co-occurrence. The CompareACE score [9], based on the Pearson correlation coefficient between the nucleotide base frequencies of two motif alignments, is used to measure the 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.

### 3.6 Mining Co-occurrences of Sites

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 homologues and over-represented oligonucleotides [10, 11]. 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 a considered gene group, and the found site co-occurrences is also called site combinations, which are with both a support value and a confidence value. In the system, a user can specify the minimum support value, the minimum confidence value and the maximum number of sites in a site combination.

In the step, the RgS-Miner detect the site occurrence associations from the combinations of the TF binding site homologues, over-represented repetitive sequences, and DNA motifs by implemented algorithm Apriori and AprioriTid algorithm [12]. The site co-occurrence detection detects co-occurring site combinations in the upstream regions. The cumulative hyper-geometric probability has been used to assess the functional significance of computationally derived motifs [13, 9, 14]. A motif pair is considered to co-occur significantly if the hyper-geometric P-value is less than the reciprocal of the total number of motif pairs tested; that is, if  $P(C>c') < 1/MP$ , where MP is the total number of site pairs considered in the analysis.

### 3.7 Interfaces

The system requires users to input the case name and description to enable the result of the analysis of each user's input case to be stored. The result in text format contains the consensus pattern of TF binding sites, the number of the upstream sequences in which the TF binding sites occur, and the description of the TF binding sites. RgS-Miner also provides the detailed information about site occurrences in the upstream regions, obtained by clicking on links on the web pages. The interface shows the over-represented (OR) oligonucleotides (also called repeats) and their corresponding p-values, which measure the over-representation of the oligonucleotides.

The system has two output pages to present site combinations - a tree-like view and a circular synergy map - to elucidate the relationship among site combinations. The tree-like view presents the site combinations that contain the pattern either on the right or left. The circular synergy map shows the synergism between putative regulatory sites.

## 4 Implementation

The biological data warehouse is implemented by using the MySQL relational database management system version 3.23, which is running on a PC server under the Linux Red Hat 8.0 operating system. The wrapper and monitors are written by C/C++ programming language. 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 from 4 to 25 bps. For the sake of efficiently providing the analysis requirements in RgS-Miner system, the index of whole genome sequences are implemented. The query forms and output pages on the web are created dynamically by CGI scripts written in PHP programming language, which accesses the database via the PHP-MySQL module.

## 5 Related Works

Many experimentally identified TF binding sites have been collected in TRANSFAC [3], which is the most complete and well maintained database of transcription factors, their genomic binding sites and DNA-binding profiles. van Helden *et al.* systematically searched promoter regions of potentially co-regulated genes for over-represented repetitive oligonucleotides, which could perhaps be transcription factor binding sites, involved in regulating genes [15]. Numerous methods including Consensus [16], MEME [8], Gibbs Sampler [7] and ANN-Spec [17], for multiple local alignment have been employed to tackle the problem for identifying individual TF binding site patterns. In many cases in which binding sites for transcription factors are experimentally determined. Brázma *et al.* [18] developed a general software tool to find and analyze combinations of transcription factor binding sites that occur in upstream regions of the genes in the yeast genome. Their tool can identify all the combinations of sites that satisfy the given parameters with respect to a given set of gene in promoter regions, their counter sets and the chosen set of sites.

RSA tools [6] is a web site for performing computational analysis of regulatory sequences, focusing on yeast. A series of computer programs have been developed and integrated for the analyzing transcriptional regulatory sequences. The TOUCAN system is a Java application for predicting cis-regulatory elements from a set of co-expressed or co-regulated genes [19]. Putative sites of known transcription factors are detected using a database of known site and a probability background model.

## 6 Discussions

The system facilitates the analysis of gene transcription regulatory sequences in a set of potentially co-regulated genes. Many computational and statistical methods have been previously developed to compare the gene expression profiles obtained from microarray experiments or cDNA libraries, and thus elucidate the gene co-expression.

DNA motif discovery methods represent the occurrence sequences of a pattern as a consensus pattern and a position weight matrix (PWM). The system currently

supports only consensus patterns in the analyses, but the PWM will be supported in the future. Limited by computational power, the system is currently restricted in its application to up to 25 genes in a submitted group, and each the length of the sequences may not exceed 3000 bps.

Comparing the predictions that pertain to multiple gene groups in various biological considerations is interesting and important work. Statistical and computational methods are designed to examine the group-specificity of the putative regulatory sites as well as the co-occurrence of regulatory sites. The authors plan to support the development of functionality to assess the group-specificity of the putative regulatory sites and co-occurrence sites in the future.

We present a novel system to integrate different approaches for detecting putative transcriptional regulatory sites in the upstream regions of genes. The system as several characteristics addressed as follows: (i) The required gene annotation information, human genome sequences, and upstream features are *prior* storing in the RgS-Miner database which is comprised in the proposed biological data warehouse. (ii) The sequences of the gene upstream regions can be tailored easily by just specifying the start and end positions. (iii) Cross-species comparison of known TF binding sites between human and other vertebrates is also provided. (iv) To efficiently obtain the background frequency of an oligonucleotide, we establish a special data structure to index whole genomic sequences. (v) To merge the redundant motifs predicted by different methods, we propose a motif grouping methodology to result non-redundant motif groups. (vi) Investigating the synergism of putative regulatory sites, the system implements a data mining algorithm to detect the site co-occurrences, and adapt statistical methods to filter insignificant site combinations. In addition to the tree-like view and the circular synergy map facilitates the representation of the regulatory site synergisms.

### Acknowledgements

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC93-3112-B-008-002. Prof. Ueng-Cheng Yang and Prof. Yu-Chung Chang are appreciated for their valuable discussions regarding molecular biology. We also thank Prof. Cheng-Yan Kao for his valuable suggestions and comments.

### References

1. Pruitt, K.D. and Maglott, D.R.: RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*, Vol. 29. 1 (2001) 137-140
2. Hubbard, T., et al.: The Ensembl genome database project. *Nucleic Acids Res*, Vol. 30. 1 (2002) 38-41
3. Wingender, E., et al.: The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*, Vol. 29. 1 (2001) 281-283
4. Ohler, U. and Niemann, H.: Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet*, Vol. 17. 2 (2001) 56-60

5. Benson, G.: Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, Vol. 27. 2 (1999) 573-580
6. Van Helden, J., et al.: A web site for the computational analysis of yeast regulatory sequences. *Yeast*, Vol. 16. 2 (2000) 177-187
7. Lawrence, C.E., et al.: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, Vol. 262. 5131 (1993) 208-214
8. Bailey, T.L. and Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, Vol. 2. (1994) 28-36
9. Hughes, J.D., et al.: Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol*, Vol. 296. 5 (2000) 1205-1214
10. Horng, J.T., et al.: Mining putative regulatory elements in promoter regions of *Saccharomyces cerevisiae*. *In Silico Biol*, Vol. 2. 3 (2002) 263-273
11. Horng, J.T., et al.: The repetitive sequence database and mining putative regulatory elements in gene promoter regions. *J Comput Biol*, Vol. 9. 4 (2002) 621-640
12. Srikant, R., et al.: Mining Generalized Association Rules., Vol. (1995) 407-419
13. Jensen, L.J. and Knudsen, S.: Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics*, Vol. 16. 4 (2000) 326-333
14. Sudarsanam, P., et al.: Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Res*, Vol. 12. 11 (2002) 1723-1731
15. Van Helden, J., et al.: Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol*, Vol. 281. 5 (1998) 827-842
16. Hertz, G.Z. and Stormo, G.D.: Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, Vol. 15. 7-8 (1999) 563-577
17. Workman, C.T. and Stormo, G.D.: ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput*, Vol. (2000) 467-478
18. Brazma, A., et al.: Data mining for regulatory elements in yeast genome. *Proc Int Conf Intell Syst Mol Biol*, Vol. 5. (1997) 65-74
19. Aerts, S., et al.: Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*, Vol. 31. 6 (2003) 1753-1764