Published online March 29, 2004

1948–1956 *Nucleic Acids Research, 2004, Vol. 32, No. 6*
DOI: 10.1093/nar/gkh345

Identifying transcriptional regulatory sites in the human genome using an integrated system

Hsien-Da Huang, Jorng-Tzong Horng[1,2,*], Yi-Ming Sun[1], Ann-Ping Tsou[3] and Shir-Ly Huang[2]

Department of Biological Science and Technology and Institute of Bioinformatics, National Chiao-Tung University, Hsin-Chu 300, Taiwan, [1]Department of Computer Science and Information Engineering, National Central University, Chung-Li 320, Taiwan, [2]Department of Life Science, National Central University, Chung-Li 320, Taiwan and [3]Insititute of Biotechnology in Medicine, National Yang-Ming University, Taipei 112, Taiwan

Received November 17, 2003; Revised January 1, 2004; Accepted March 1, 2004

ABSTRACT

This work develops an integrated system which, after a set of genes are inputted, is able to predict transcriptional regulatory sites and to detect the co-occurrence of these regulatory sites. The system integrates several site detection methods such as known site matching, over-presented oligonucleotide detection and DNA motif discovery programs. User profiles and history pages enable users to trace the sequence analyses of these transcriptional regulatory sites. Two groups of co-regulated genes were used to test the proposed system. The results predicted by the proposed system consist of known site homologs and putative regulatory sites. By comparing these sites with previously published results, the proposed system is able to help biologists identify possible candidates for the regulatory sites from groups of co-regulated genes. The integrated system is now available at http://rgsminer.csie.ncu.edu.tw/.

INTRODUCTION

Genome-wide gene expression data provide a unique dataset of genes and these are used to decipher the mechanisms that underlie the common regulation of the transcriptional response. Gene regulation is one of the most challenging and exciting areas in molecular genetics. The large amount of information gained from the various projects involved in the sequencing of the human genome and the elucidation of gene expression within the human genome enables researchers to use a computational approach to investigating the mechanism by which genes are regulated. Not only does the identification of transcription factor (TF) binding sites yield valuable information on gene expression and regulation, but also the detection of the co-occurrence of regulatory sites facilitates analysis of regulation mechanisms. Recently, biological information and analytical methods have become available for analyzing gene expression and transcriptional regulatory sequences. However, users must set up complicated analyses that query the data in various databases, and then they must analyze the regions upstream of the genes using various predictive tools before finally converting the results between formats. Beyond methods for predicting transcriptional regulatory sites, new automated and integrated methods for analyzing gene upstream sequences at a higher level are urgently required. Identifying regulatory sites requires a large number of biological databases, so efficient and integrated data management methods are essential.

Many experimentally identified TF binding sites have been collected in TRANSFAC (1), which is the most complete and well maintained database of TFs, their genomic binding sites and DNA-binding profiles. Van Helden *et al.* (2) systematically searched promoter regions of potentially co-regulated genes for over-represented repetitive oligonucleotides, which might perhaps be TF binding sites and involved in regulating genes (2). Their work applied a simple and fast method for the identification of isolating DNA binding sites for TFs within families of co-regulated genes. They presented results for *Saccharomyces cerevisiae*.

Numerous methods including Consensus (3), MEME (4), Gibbs Sampler (5) and ANN-Spec (6), for multiple local alignment have been employed to tackle the problem of the identification of individual TF binding site patterns. In many cases where the binding sites for TFs have been experimentally determined, software has been shown to yield the known binding site patterns, indicating that such methods can discover unknown TF binding sites from a collection of sequences believed to contain a common binding site pattern.

Brazma *et al.* (7) developed a general software tool to find and analyze combinations of TF binding sites that occur in the upstream regions of genes in the yeast genome. As well as analyzing the association rules in the combinations, they analyzed the appearances of the combinations in the promoter and in regions that were randomly chosen. Their tool can identify all the combinations of sites that satisfy given parameters with respect to a given set of genes in promoter

*To whom correspondence should be addressed at Department of Computer Science and Information Engineering, National Central University, Chung-Li 320, Taiwan. Tel: +886 3 4227151; Fax: +886 3 4273485; Email: horng@db.csie.ncu.edu.tw

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Table 1. Sources of genes and upstream features in the database

Category	Entries	Source	URL	References
Human genome sequence	33 840	GenBank at NCBI	http://www.ncbi.nlm.nih.gov/	(26)
Genes	24 847	Ensembl	http://www.ensebml.org/	(27)
TFs and binding sites	16 311	TRANSFAC	http://www.gene-regulation.com/pub/databases.html	(1)
TSSs	60 540	Ensembl and Eponine	http://www.ensembl.org/; http://www.sanger.ac.uk/Users/td2/eponine/	(27,28)
Tandem repeats	704 788	Ensembl and Tandem-Repeat-Finder	http://www.ensembl.org/; http://c3.biomath.mssm.edu/trf.html	(27,29)
Repeats	5 566 532	HGB and Repeat-Masker	http://genom,e.ucsc.edu/; http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker	(30,31)
CpG islands	9374	HGB	http://genome.ucsc.edu/	(31)

regions, their counter-sets and the chosen set of sites. Horng and coworkers (8,9) also predicted putative regulatory sites by detecting combinations of known site homologs and over-represented (OR) oligonucleotides in the upstream regions of the genes in yeast genome by using a data mining approach.

The type of analysis performed in the investigation of the gene transcriptional regulations involves three processes: generating gene expression profiles, analyzing transcriptional regulatory sequences and detecting co-occurrence sites.

RSA-tools (10) is a website for performing computational analysis of regulatory sequences, and focuses on yeast. A series of computer programs have been developed and integrated for analyzing the transcriptional regulatory sequences. RSA-tools has recently been extended to other organisms such as archaea, bacteria and *Homo sapiens*. The TOUCAN system is a Java application for predicting *cis*-regulatory elements from a set of co-expressed or co-regulated genes (11). Putative sites of known TFs are detected using a database of known site and a probability background model. However, TOUCAN does not allow the detection of the co-occurrence of regulatory sites.

An integrated system for analyzing transcriptional regulatory sites in the human genome was designed and implemented. The analytical results in each phase are stored in the database. Users can input a gene group or a set of upstream sequences and then work stepwise on the analysis of the transcriptional regulatory sequences. 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 and is able to predict transcriptional regulatory sites. The predictive system facilitates the detection of regulatory sites in upstream regions of the genes and will help the discovery of co-occurrence among the regulatory sites. Two groups of co-regulated genes were used to test the proposed system. The results predicted by the system are made up of known site homologs as well as putative regulatory sites. By comparing these sites with previously published results, the proposed system should be able to help biologists identify possible targets acting as regulatory sites from sets of co-regulated genes.

MATERIALS AND METHODS

The system is designed to record information about human genomic sequences, TFs and TF binding sites, gene transcriptional start sites (TSSs), repetitive elements and CpG islands in a database, as presented in Table 1. The system facilitates the detection of regulatory sites in the human genome. The gene group to be used may be constructed based on cluster analysis of gene expression data or from the genes considered potentially to be co-regulated under particular transcriptional regulation mechanisms; once identified, this group is inputted to the system. Graphical web interfaces are used to show the upstream regions in which regulatory sites or combinations thereof are identified. The database also stores user profiles and analytical histories.

System flow

Figure 1 shows the system flow for analyzing transcriptional regulatory sequences. Users first input a set of genes or a set of upstream sequences, which can be constructed from various gene expression analyses. Before predicting regulatory sites in upstream regions, the system enables users to query the gene annotations and sequences in the database, and to tailor the upstream sequences to specified regions of the genes. The preprocessing phase thus 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 used to predict regulatory sites. Users run each predictive method separately to detect the regulatory sites in the upstream regions. As shown in Figure 1, a user can search the known TF binding sites within the submitted upstream regions to determine the existence of known site homologs. Unlike the matching to known TF sites, OR oligonucleotides in the upstream regions are detected by comparing the number of occurrences of sites in the upstream regions with the background frequency, rather than by searching the TF binding site database. DNA motif discovery methods are then employed to detect statistically any putative regulatory sites in the upstream regions. Many highly similar regulatory motifs are thus detected. The system then has a function that groups the redundant motifs and 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 upstream regions, a statistical measurement, 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. In addition, during the analysis users can annotate their input cases. The results of each step of the analysis are automatically stored in the

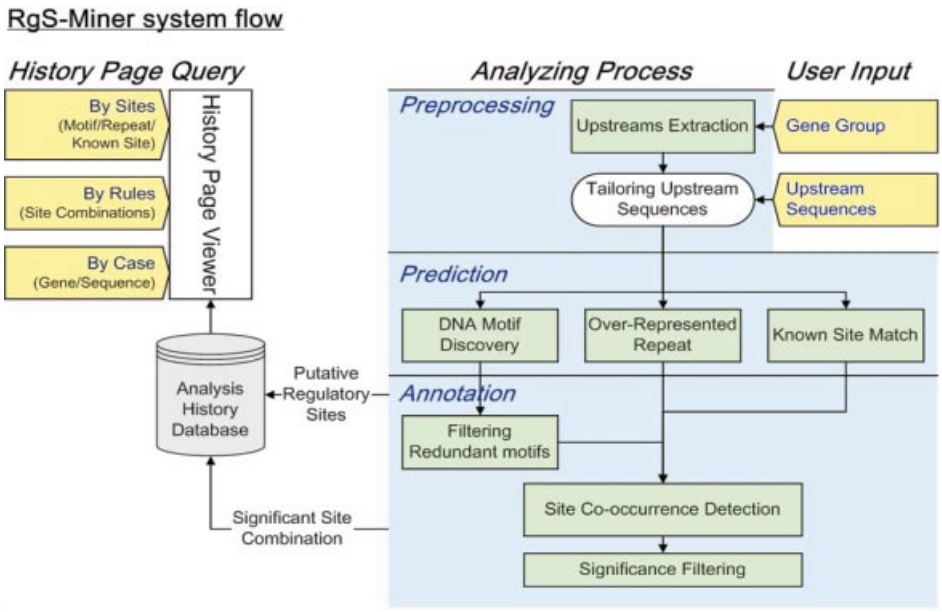

Figure 1. System flow.

database. Users can log in the system to query user profiles and history pages, which are then displayed on the web pages. The history pages intuitively present the annotations and analytical results, including putative regulatory sites and statistically significant combinations of co-occurring sites.

Preprocessing

The gene upstream regions can be obtained from the database using a query or from user submitted sequences, if the gene instances are not found in the database. The length of the upstream sequence is restricted to 3000 bp before the gene start position. The positions at which the gene coding begins, and the predicted TSSs, are obtained from the Ensembl database. Some genes within this database do not have the annotations for a predicted TSS; in this case users can tailor the upstream regions by referring to the start positions of the gene coding regions.

Predicting the regulatory site

Oligonucleotide analysis. The oligo-analysis has been developed to detect OR oligonucleotides in upstream regions. It is based on a systematic counting of occurrences of all possible oligonucleotides in a given sequence (10,12). An advantage of the method is that it can detect all the over-represented patterns for a given length in a single run. The system applies a statistical method to discover statistically significant oligonucleotides. These are DNA sequences of small length within the upstream regions of genes that are identified by comparing their frequencies of occurrence in the regions to their background frequencies of occurrence throughout human genome. The frequencies of occurrence of oligonucleotides produce a preconstructed index.

Another significant measure of oligonucleotides (12) is the frequency of occurrence of an oligonucleotide within upstream regions in relation to that of all human non-coding sequences as a background. Here, OR oligonucleotides in the upstream regions of selected genes are detected. If $F_e(b)$ is the occurrence probability of oligonucleotide b in all non-coding regions of the human genomic sequence, then the b oligonucleotide would also be expected to occur $u = T \times F_e(b)$ times in the upstream regions of genes, where T represents the total number of possible matching positions of an oligonucleotide of length w across both strands of the sequence set. Using a simple binomial model, the standard deviation of oligonucleotide occurrences becomes $\sigma = \{T \times F_e(b) \times [1 - F_e(b)]\}^{1/2}$. Let n be the frequency of the considered oligonucleotide b occurring in the upstream regions; the Z-score is given by $Z = (n - u)/\sigma$. The probability of observing at least n successes, as given by Chebyshev's theorem, is less than or equal to $p = [(n - u)/\sigma]^{-2}$. If $Z > 0$, then a lower p-value corresponds to a more over-represented oligonucleotide. If $Z < 0$, then a lower p-value corresponds to an under-represented oligonucleotide.

TRANSFAC known site matching. The experimentally identified TF binding sites were obtained from TRANSFAC (professional 5.4), which contains 11 537 sites and 4774 factors (1). In the system, 3294 vertebral TF binding sites are matched to upstream regions 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 use of a mismatch penalty. The known TF binding sites are matched to the prepared upstream regions in both strands; the positions of each known site homolog are then stored in the database for further analysis during the annotation phase.

DNA motif discovery: Gibbs Sampler, MEME and AlignACE. Three popular regulatory site prediction programs (Gibbs Sampler, MEME and AlignACE) were integrated into this program to aid discovery of DNA motifs and thus identify the binding sites in a group of upstream regions. The Gibbs Sampler [Lawrence's Gibbs Motif Sampler Version 1.01.009 (5)] is used as the option 'site sampler'. One hundred 'seeds'

or starting points are used, a maximum of 2000 iterations are performed for each run and the highest scores are reported. The MEME algorithm uses an expectation maximization algorithm to find patterns in the input sequences. MEME version 2.2 (4) is built into our system and the highest scores for this approach are reported. AlignACE (13) is based on a Gibbs sampling algorithm and returns a series of motifs that are over-represented in the upstream regions of the inputted genes. Then numerous differences between AlignACE and Gibbs Sampler have been addressed previously (13). The motifs obtained by the DNA motif discovery methods are stored in the form of consensus patterns and this includes the site sequences that occur in each upstream region.

Filtering redundant regulatory motifs

It is clear that some of the DNA motifs detected by various approaches will be highly similar to each other and are thus redundant as far as further analysis for detecting site co-occurrence. The CompareACE score (13), 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 similarity between each pair of motifs is then clustered. 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 group nearest to the centroid of the motif cluster is finally selected as the representative motif of the motif group.

Detecting co-occurrences of sites

A previous study of regulatory site prediction by Horng and coworkers (8,9) used a data mining method to identify the associations between site occurrences with combinations of known TF binding site homologs and OR oligonucleotides. That method is herein extended to three categories of potentially regulatory sequences, i.e. known site homologs, OR oligonucleotides and DNA motif groups. Accordingly, the implemented algorithm detects sites that occur concurrently in the upstream regions of a considered gene group, and the site co-occurrences found are called site combinations which are given both a support value and a confidence value. A combination of sites is defined in two parts, the left and the right parts. For example, a site combination '[Known Site] ACGCCC $\rightarrow$ [Repeats] CACGCC' indicates that a known site homolog ACGCCC and an OR oligonucleotide CACGCC occur concurrently within a given set of upstream sequences of co-expressed genes. 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.

Statistical considerations associated with site co-occurrence

The site co-occurrence detection detects co-occurring site combinations in the upstream regions. Cumulative hypergeometric distribution is incorporated to filter out insignificant site combinations. The basic idea is that the sites to the left and to the right of a site combination may emerge independently in the upstream regions of a gene group. Thus a combination of sites is divided into left and right parts represented, for example, as 'aaatat, ttgaa $\Rightarrow$ gcggag'.

The cumulative hypergeometric probability is used to assess the functional significance of computationally derived motifs (13–15). A motif pair is considered to co-occur significantly if the hypergeometric p-value is less than the reciprocal of the total number of motif pairs tested.

Implementation

The proposed system is implemented using the MySQL relational database management system version 3.23, which runs on a PC server under the Linux Red Hat 8.0 operating system. The query forms and output pages on the web are generated dynamically using CGI scripts written in PHP programming language which accesses the database via a PHP-MySQL module. An index of human genome sequences is established to return quickly and efficiently the occurrences of a query pattern, a short nucleotide sequence (4–25 bp) within the human genome. Huge computations are inevitable when oligonucleotide analysis is conducted in order to identify regulatory sites in the upstream regions of genes (10,12), so a highly efficient strategy is required when dealing with the very large size of the human genome. The system described here implements an index constructed from the human genome sequences, and this reduces the algorithmic complexity of a search for an oligonucleotide in a human genome sequence. In this study, the lengths of the oligonucleotides of regulatory sites did not normally exceed 25 bp. The proposed system supports a query for the number of occurrences of oligonucleotides whose lengths are between 4 and 25 bp. In an oligonucleotide analysis conducted to discover OR oligonucleotides in the upstream regions of genes, the index efficiently returned the occurrence frequencies of all the oligonucleotides with lengths from 4 to 25 bp. The index is designed efficiently to re-index the entire genome sequences whenever genome assembly sequences are updated.

RESULTS

Data input pages

Before transcriptional regulatory sites can be predicted and analyzed, a set of genes with accession numbers or gene names is inputted into the proposed system. In addition, while submitting these genes, the available fields are the start position of the gene coding region, the end position of the coding region, the boundaries of the exon regions, predicted information about the TSSs, the locations of CpG islands and the occurrences of the repetitive elements in the gene upstream regions. Users can tailor the upstream sequences to their requirements by adjusting the start positions of the gene coding regions or the TSSs. Furthermore, users can input a set of upstream nucleotide sequences for analysis of transcriptional regulatory sites. 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. For instance, a case called 'known cell cycle regulated genes G2/M′ contains the known regulated genes in the G2/M phase.

Format of outputs

The system can present the analyzed results in various output formats. Known TF binding sites are identified in the upstream

sequences and this is prepared in the preprocessing phase. The result in text format contains the consensus pattern of the TF binding sites, the number of upstream sequences in which the TF binding sites occur and the description of the TF binding sites. For example, the consensus pattern ACGCCC of the known TF binding site with the identifier HS$U2SN and TRANSFAC accession number R01498 occurs in 10 of 13 upstream regions. The system also provides detailed information about site occurrences in the upstream regions, obtained by clicking on links on the web pages. The interface shows the OR oligonucleotides (also called repeats) and their corresponding *p*-values, which measure the over-representation of the oligonucleotides. The system cumulates not only the occurrences of an oligonucleotide in the considered upstream regions, but also the background occurrences in both coding and non-coding regions of the human genome, upon receiving an *i*-Human query. *i*-Human indexes the entire human genomic sequence, chromosome by chromosome, and efficiently returns the occurrence positions of an oligonucleotide. For example, the *p*-value of the OR oligonucleotide caccg is 0.00757, and this oligonucleotide occurs 19 times in eight of the 13 submitted upstream regions. The background frequency of the oligonucleotide is 1 243 432 throughout the genome, and is 576 823 in coding regions.

A motif group contains at least one motif predicted by DNA motif discovery methods. For instance, the AlignACE pattern GBKSCCYRGR is the representative motif in a motif group, which contains other motifs such as GRGBGCRGKG from AlignACE and the pattern CTGGGYAACA identified using the Gibbs Sampler.

Furthermore, the system detects the co-occurrence of putative regulatory sites including known site homologs, OR oligonucleotides and DNA motif groups. The output page displays significant site combinations with chi-square values, *p*-values, support values, confidence values and the number of occurrences in the upstream regions considered. 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 of the sites in the right parts of a combination. The default thresholds of chi-square value and *p*-value are set to 3.84 and 0.001, respectively. The support value of a site combination is the percentage of the upstream regions that contain the sites. The confidence value is the percentage of upstream regions that contain the site in the left part and that also contain the site in the right part. The default support and confidence thresholds are set to 0.5 and 0.6, respectively. For example, a site combination '[Known Site] ACGCCC → [Known Site] CACGCC' has a *p*-value of 0.035 (<0.05). The support and confidence values are 0.62 and 1.0, respectively. The positions of occurrences of combinations are depicted graphically on the output pages.

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 on either the right or the left. The circular synergy map shows the synergism between putative regulatory sites; the map includes a line between any pair of sites on the circumference if these two sites are found in a site combination. The circular synergy map is a dynamic web page.

User profiles and case histories

The system described here maintains user profiles and histories of their analyses of transcriptional regulatory sites. All information, including the analyzing parameters and the results, are stored in the database to be queried subsequently. The user profile is also recorded and all the available private and public cases are listed on the web page. The page for querying histories presents the case name, the case description, the source sequences in the gene upstream regions, putative regulatory sites and site combinations. The information allows users to trace all analyzed cases. A tree-like presentation allows all the results of a case to be displayed on a single page. Accordingly, the history pages show four main categories of information: case description, source sequences, site and motif groups and site combinations.

Case study 1: Tissue-specific differentially expressed genes
Gene groups: Tissue-specific differentially expressed genes
Gene expression analysis: cDNA libraries from ovary, liver, uterus, brain, retina and skeletal muscle
Gene expression analysis: *R*-statistics and PAC values
Organism: *Homo sapiens*
Regulatory site prediction methods: Oligo-analysis and known site matching
Detection of site co-occurrence: Yes

Very-large-scale gene expression data, as in the UniGene and dbEST database, are able to find genes with markedly significant differential expression in specific tissues. The differentially expressed genes in a specific tissue are potentially regulated concurrently by a combination of TFs. This work attempts to examine the upstream regions of a group of differentially expressed genes to predict TF binding sites and to determine the way in which combinations of the known site homologs, OR oligomers and DNA motifs are distributed in the upstream regions. In a previous study, some of the authors performed a computational and statistical analysis on a large set of gene expression data, pertaining to six adult human tissues: uterus, ovary, brain, liver, skeletal muscle and retina (16). *R*-statistics were obtained (17) and an AC (18) test was performed to identify the differentially expressed genes. Table 2 presents three gene groups which are differentially expressed in the ovary, liver and skeletal muscle.

Furthermore, potentially co-regulated genes, which are the differentially expressed genes discovered in specific tissues, are of particular interest. The upstream regions are from –1 bp to –2000 bp, where +1 bp indicates the position of the gene coding region. Table 1s (Supplementary Material) lists several interesting and significant site combinations, mined from genes that are differentially expressed in the tissues 'ovary', 'liver' and 'skeletal muscle'. For example, the site combination 'acaggc, CATTT, GGTTA ⇒ agctga' is discovered in the ovary data with a confidence of 1.0, a support of 0.70 and a *p*-value of 0.003 (<0.05). CATTT and GGTTA are known sites with identifiers HS$GMCSF_03 and HS$WT1_04, respectively, in TRANSFAC, and acaggc and agctga are considerably over-represented repetitive oligonucleotides. Similarly, 'TATAA ⇒ ctcagc' refers to a site combination

Table 2. Differentially expressed genes in ovary, liver and skeletal muscle tissues

Tissue type	Genes	Gene names
Ovary	10	HSPCB, EEF1G, RPS16, SPINT2, RPL7A, RPS2, PKM2, RPS3, ENO1, K-ALPHA-1
Liver	21	AZGP1, VTN, NNMT, APOH, BF, PLG, HP, AMBP, CRP, APOA1, SERPING1, ALB, TTR, GC, ALDOB, SERPINA3, APOA2, TF, C3, SERPINA1, HPX
Skeletal muscle	35	COX7C, RPL10, HSPB1, LOC54543, RPLP1, NEB, COX7A1, MYH1, TPT1, MYH7, MYH2, ACTA1, CRYAB, TNNT3, TNNT1, MYL2, TPM1, TNNI2, TNNI1, RPS4X, RPS25, ENO3, MB, TNNC1, MYL1, MYBPC1, TTN, TNNC2, MYOZ1, DES, RPL37A, TPM2, CKM, RPL37, TCAP

Table 3. Known cell cycle regulated gene groups

Phase	Gene symbol	No. of genes
G1	E2F5	1
G1/S	CCNE1, CCNE2, CDC25A, CDC45L, CDC6, CDKN3, E2F1, MCM2, MCM6, NPAT, PCNA, SLBP, CDKN1A	13
S phase	BRCA1, BRCA2, CCNG2, CDKN2C, DHFR, MSH2, NASP, RRM1, RRM2, TYMS	10
G2	CCNA2, CCNF, TOP2A, CENPF,	4
G2/M	BIRC5, BUB1B, CCNB1, CCNB2, CDC2, CDC25B, CDC25C, CENPA, CKS1, CKS2, PLK, CDKN2D, RACGAP1, RAB6KIFL	14

in the liver with a confidence value, a support value and a *p*-value of 0.67, 0.94 and 0.011, respectively.

Table 2s (Supplementary Material) presents examples of occurrences of the association 'aacaag, HS$GG_22 => HS$GRH_03' in the tissue 'liver'. For example, the fourth row in Table 2s presents the association 'aacaag**,** HS$GG_22 ⇒ HS$GRH_03' mined from the gene AMBP using UniGene cluster IDs 'Hs.76177'. The '[-1999]-%cttgtt-[227]-%ATTTA-[94]-%ATTTA-[6]-TAAAT-[135]-%ATTAG-[153]-TAAAT-' gives the consequent positions of the known site homologs and the over-represented repetitive oligonucleotides within the association 'HS$GG_22/CTAAT' 'HS$GRH_03/TAAAT' or 'aacaag'. The first number, –1999 bp, indicates that the location of the site cttgtt is 1999 bp before the end of the upstream region, and that the site cttgtt is the reverse complement sequence of site aacaag. The symbol % indicates that the site is present in the anti-sense strand. The number, such as 227, specifies the distance between the first and second occurrences.

The repetitive oligonucleotides associated with the known TF binding sites are called putative regulatory sites because of the correlation between their occurrences with the known signal of transcriptional regulation in a group of genes differentially expressed in a particular tissue. For example in Table 3s (Supplementary Material), the site AAGAGG in the first row is present in 10 upstream regions from 10 genes in the ovary. A total of 26 occurrences are detected and the expected number of occurrences is 8.06. The *Z*-score is 6.32. The sites AAGAGG, AGAGGC, GGAGGA and CGGAGG are aligned to generate the consensus sequence MRGAGGM.

Notably, the putative regulatory site GASCTCC found in the tissue 'ovary' is similar to sites previously characterized using TRANSFAC. They are CACCTCC (RAT$A12COL_01), CAACT (HS$PR264_06) and GGAGC (RAT$MLC_04). The specific binding TFs of these putative regulatory sites are c-Myb, MAPF2 and YY1. Similarly, many putative regulatory sites discovered in other tissues, such as skeletal muscle, can be matched to known sites.

Case study 2: Known cell cycle regulated genes
Gene groups: Known cell cycle regulated genes
Gene expression analysis: In the literature
Organism: *Homo sapiens*
Regulatory site prediction methods: Oligo-analysis, known site matching and DNA motif discovery
Detection of site co-occurrence: Yes

The proposed system is applied to another case involving the cell cycle regulated genes in the human liver. Whitfield and coworkers established a microarray to measure the expression profile of HeLa cell cycle genes, and their results are available at http://genome-www.stanford.edu/Human-CellCycle/HeLa/. They also compiled a list of genes and determined using traditional methods that all these genes were cell cycle regulated in cells synchronized independent of serum stimulation (19). The data in Table 3 are rearranged into known cell cycle regulated gene groups in Table 1s with gene groups for the phases of G1, G1/S, S, G2 and G2/M.

These four gene groups G1/S, S phase, G2 and G2/M were then submitted to the proposed system to predict regulatory sites and detect site co-occurrences. The upstream regions of all genes were uniformly set from –1 bp to –1200 bp.

Table 4s (Supplementary Material) presents the known sites matched in the G2/M phase gene upstream regions. For example, the site sequence ACGCCC is a TRANSFAC known site with TRANSFAC identifier HS$U2SN_04. The TF Sp1 can bind to the site. The site occurs in 10 gene upstream regions out of 14 genes in the G2/M phase. Table 5s (Supplementary Material) presents the OR oligonucleotides mined from the upstream regions of the co-regulated gene in G2/M phase. For example, the oligomer gggcg occurs in 11 genes out of 14 gene upstream regions and is measured as an OR oligomer in the upstream of the co-regulated genes in whose *Z*-score is 17.86 and *p*-value is 0.0031.

After the DNA motif discovery tools, such as the Gibbs Sampler, MEME and AlignACE are applied, the redundant

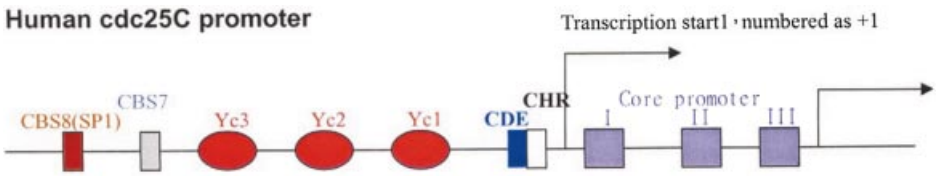

Figure 2. A verified upstream of human cdc25c.

motifs are filtered out. Table 6s (Supplementary Material) shows the motif groups discovered in the G2/M phase. For example, the first entries in Table 6s, the motifs GGBTGCRGKG and GSBKSSVGDG, are predicted by AlignACE programs. By comparing the consensus pattern with TRANSFAC known sites, it is shown that the TF Sp1 can bind to the TF binding site CACCC, and this is similar to the motifs above.

Table 7s (Supplementary Material) lists the significant site combinations, including known site homologs, OR oligonucleotides and motif groups in the G2/M phase. For example, '[Known Site] ACGCCC → [Repeats] CACGCC' in the first row of the site combination is mined from 10 genes out of 14 genes in the G2/M phase. The site to the left is present in 10 upstream regions out of 14 genes, and the OR oligonucleotide to the right is present in 11 gene upstream regions. The chi-square value and *p*-value are 9.55 and 0.011, respectively.

Figure 2 shows the CDC25C upstream regions (20) for the G2/M phase considered. For example, the human cdc25C promoter's regulatory sites include cell division control elements (CDE), the cell cycle gene homology regions (CHR), and Yc1, Yc2, Yc3, CBS7 and CBS8 regulatory elements. CHR and CHE can be predicted by matching to known sites. Yc1 and Yc2 sequences are similar to Yc3. In Table 8s (Supplementary Material), the OR oligonucleotides gcgag and cacgc are matched to the sites just characterized. In this case study, the predictions of RgS-Miner are verified by comparison with the literature. The known site homologs in the input upstream sequences and the statistically significant DNA motifs, which are putative TF binding sites, can be easily detected.

DISCUSSION

This work presents an integrated method for the detection of putative transcriptional regulatory sites and their co-occurrences in the upstream regions of genes. 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 developed to compare the gene expression profiles obtained from microarray experiments or cDNA libraries, and thus elucidate gene co-expression. The analysis of gene expression profiles has been addressed in relation to various types of genes including, for example, cell cycle co-regulated genes and differentially expressed genes. The system described herein focuses on the designation of gene groups or gene upstream sequences as the inputs for predicting regulatory sites and supporting further investigation. The biological interpretation of putative regulatory sites and site co-occurrences is considered to be the endpoint of the submitted gene groups. For example, a gene group of liver-specific genes are co-regulated in the G2/M phase. The putative regulatory sites or site co-occurrences mined from the group of genes are thus considered potentially to be involved in inducing or repressing the transcription of genes.

Utilization of the TF database TRANFAC (21,22) to search all matching positions within a given set of upstream sequences is a common practice in gene regulation study (7,11,23–25). The odds of finding that some of the matched positions are not bona fide binding sites is inevitable when the homology search was targeted at short oligomers of 5–10 bp. This outcome is likely to become more serious when the search is conducted for the upstream sequences of a single gene rather than a set of co-expressed genes. The system we reported here, RgS-Miner, does not assume that binding sites are hidden in a noisy background sequence, unlike the MotifScanner in the Toucan system (11).

In RgS-Miner, not only do we match all the data entries in TRANFAC within a set of gene upstream regions, but we also compute the occurrence of the matched known site within a set of co-regulated gene upstream regions. To obtain the occurrence of the matched known site within a set of co-regulated gene upstream regions, the system computes the fraction of gene upstream regions that contain a known site after matching all the TRANSFAC sites to each gene upstream region of a co-expressed set. Known sites occurring in less than half of the set of gene upstream regions (set default to 50% of gene upstream regions) are removed. The *p*-value computed based on a cumulative hypergeometry distribution is used to assess the significance of the site co-occurrences. If a known site is correlated to other site sequences, which can be known sites, OR oligonucleotides or DNA motifs, the system has more confidence that the known site may be a bona fide binding site.

It should be noted that the TRANSFAC matching results in the prediction phase are stored. Also, the statistically significant site co-occurrences found in the annotation phase are stored. These two results related to TRANSFAC under different conditions are provided by the system and can be refined by users.

RSA-tools is a well developed system for analyzing transcriptional regulatory sequences, with a special focus on yeast (10). It has now been extended to other organisms such as archaea, bacteria and *Homo sapiens*. Table 4 compares the proposed system with RSA-tools. The proposed system focuses only on the human genome, whereas RSA-tools supports the human genome and over 100 other organisms. A gene group or a set of gene upstream sequences can be input to both systems, which support the tailoring of the specific upstream sequence regions. The main differences between the TOUCAN system (11) and the RgS-Miner system are as

Table 4. Comparison of RSA-tools, the TOUCAN system and the RgS-Miner system

Comparison items	RSA-tools (10)	TOUCAN system (11)	Proposed system
Support organisms	112 organisms (archaea, bacteria, yeast and human)	Support organisms in Ensembl (27)	Human genome
Known site data sources	User inputted	User inputted	TRANSFAC
Input data type (gene names or gene upstream sequences)	Both	Query from Ensembl	Both
Analysis of regulatory sequences			
Oligonucleotide analysis	Oligo-analysis (2), dyad-analysis (32)	Oligo-analysis (2)	Oligo-analysis (2), *Z*-score (12)
Known site match	Known site submitted as query patterns	MotifScanner (11)	Yes
DNA motif discovery	Gibbs Sampler (5), Consensus (3)	MotifSampler (11) (Gibbs sampling)	Gibbs Sampler (5), AlignACE (13), MEME (4)
Providing pattern background frequencies	Yes	No	*i*-Human (the index of human genomic sequences)
Filtering or comparing redundant motifs	Not supported	Not supported	Yes
Pattern matching	Yes	Not supported	Not supported
Detection of site co-occurrence	Not supported	Not supported	Mining association-rule algorithm (8)
Filtering insignificant site co-occurrence	Not necessary	Not supported	Chi-square test and cumulative hypergeometric probability
Feature view	Yes	Yes	Yes
Circular synergy map	Not necessary	No	Yes
System properties			
Database management system	No	Supported by Ensembl	Yes
Interface	Web-based	Java application	Web-based
History pages	Not supported	Not supported	Yes
User profiles	Not supported	Not supported	Yes
Systematic analysis	Medium	High	High
Gene annotation reference links	None	Yes	Yes

follows. First, the TOUCAN system applies MotifScanner to search for instances of known motifs by considering background models. The RgS-Miner system incorporates a pattern match module for finding instances of TRANSFAC known motifs (consensus sequences) without using statistically addressed background models. Secondly, the TOUCAN system employs the Gibbs Sampler to identify new patterns of regulatory motifs. In RgS-Miner, the Gibbs Sampler, MEME and AlignACE are used separately to find DNA motifs. The proposed system includes a module that groups similar motifs into sets to delete the redundant DNA motifs predicted by various approaches. Finally, the RgS-Miner system facilitates the detection of the site co-occurrences within a putative sequence set. Details of the method are presented elsewhere (8,9,16). The TOUCAN system does not detect site co-occurrences.

The differences between the implementation criterion of the TOUCAN system and of the RgS-Miner system are as follows. First, TOUCAN is a Java application. RgS-Miner performs a series of analyses on the website. Secondly, the RgS-Miner system stores all user profiles and the analytical results of each case in a database. Finally, the RgS-Miner system includes a data warehouse that stores and maintains heterogeneous biological databases. Other biological databases can be more easily incorporated into the RgS-Miner system than into the other system.

The system includes a database populated from TRANSFAC to allow the matching of known TF sites in gene upstream regions; RSA-tools does not include such a database. In contrast, the pattern-matching feature in RSA-tools enables users to search for known site homologs in the submitted sequences obtained by inputting a TRANSFAC consensus pattern. RSA-tools provides two DNA motif discovery methods: the Gibbs Sampler (5) and Consensus (3). However, the RgS-Miner system includes three methods: the Gibbs Sampler, MEME (4) and AlignACE (13). Additionally, the system prunes redundant motifs into motif groups and implements a data mining algorithm to determine the site co-occurrences. Insignificant site combinations are filtered out by a statistical method, the cumulative hypergeometric distribution. A graphical visualization interface to display the co-occurrence of sites in combinations is implemented as a circular synergy map. Notably, RSA-tools does not detect site co-occurrences. In particular, the RgS-Miner system has excellent functionality that facilitates the management of analyses of transcriptional regulatory sequences, including historical pages, user profiles and other databases.

The system can analyze transcriptional regulatory sequences for an input gene group after the gene group is inputted. The modules in the system are then executed, including those for matching known sites, detecting OR oligomers, discovering DNA motifs and identifying site co-occurrence. The results of the analysis are analyzed for statistical significance. Although the biological results of the *in silico* analyses are promising, the predictions of transcriptional regulatory sequences need to be further tested by biological investigations.

DNA motif discovery methods can involve the occurrence of sequences as a pattern to form a consensus pattern as well as an alternative, the position weight matrix (PWM) method. The system currently supports only consensus patterns in the analyses, but the PWM will be supported in the future. Because there are limitations on the computational power available, the system is currently restricted to $\leqslant$25 genes as the submitted group, and the length of these sequences may not be >3000 bp.

Comparing the predictions that pertain to multiple gene groups in various biological considerations is interesting and very important. 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 assist the identification of group-specific putative regulatory sites and co-occurrence sites in the future.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors appreciate the valuable contributions of Professor Ueng-Cheng Yang and Professor Yu-Chung Chang regarding molecular biology. They also thank Professor Cheng-Yan Kao for his valuable suggestions and comments. The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 92–3112-B-008–003.

REFERENCES

1. Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhauser,R. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.
2. van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.
3. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
4. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
5. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
6. Workman,C.T. and Stormo,G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.*, 467–478.
7. Brazma,A., Vilo,J., Ukkonen,E. and Valtonen,K. (1997) Data mining for regulatory elements in yeast genome. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 65–74.
8. Horng,J.T., Huang,H.D., Huang,S.L., Yan,U.C. and Chang,Y.C. (2002) Mining putative regulatory elements in promoter regions of *Saccharomyces cerevisiae*. *In Silico Biol.*, **2**, 263–273.
9. Horng,J.T., Huang,H.D., Jin,M.H., Wu,L.C. and Huang,S.L. (2002) The repetitive sequence database and mining putative regulatory elements in gene promoter regions. *J. Comput. Biol.*, **9**, 621–640.
10. van Helden,J., Andre,B. and Collado-Vides,J. (2000) A web site for the computational analysis of yeast regulatory sequences. *Yeast*, **16**, 177–187.
11. Aerts,S., Thijs,G., Coessens,B., Staes,M., Moreau,Y. and De Moor,B. (2003) Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res.*, **31**, 1753–1764.
12. Levy,S., Hannenhalli,S. and Workman,C. (2001) Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics*, **17**, 871–877.
13. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.
14. Jensen,L.J. and Knudsen,S. (2000) Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics*, **16**, 326–333.
15. Sudarsanam,P., Pilpel,Y. and Church,G.M. (2002) Genome-wide co-occurrence of promoter elements reveals a *cis*-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Res.*, **12**, 1723–1731.
16. Huang,H.D., Chang,H.L., Tsou,T.S., Liu,B.J., Kao,C.Y. and Horng,J.T. (2003) *Third IEEE Symposium on BioInformatics and BioEngineering*. Computer Society, IEEE, Bethesda, MD, pp. 297–304.
17. Stekel,D.J., Git,Y. and Falciani,F. (2000) The comparison of gene expression from multiple cDNA libraries. *Genome Res.*, **10**, 2055–2061.
18. Audic,S. and Claverie,J.M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.
19. Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.
20. Korner,K. and Muller,R. (2000) *In vivo* structure of the cell cycle-regulated human cdc25C promoter. *J. Biol. Chem.*, **275**, 18676–18681.
21. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
22. Heinemeyer,T., Chen,X., Karas,H., Kel,A.E., Kel,O.V., Liebich,I., Meinhardt,T., Reuter,I., Schacherer,F. and Wingender,E. (1999) Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res.*, **27**, 318–322.
23. Brazma,A., Jonassen,I., Vilo,J. and Ukkonen,E. (1998) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, **8**, 1202–1215.
24. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
25. Sosinsky,A., Bonin,C.P., Mann,R.S. and Honig,B. (2003) Target Explorer: An automated tool for the identification of new target genes for a specified set of transcription factors. *Nucleic Acids Res.*, **31**, 3589–3592.
26. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
27. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
28. Ohler,U. and Niemann,H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.*, **17**, 56–60.
29. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
30. Jurka,J. (1998) Repeats in genomic DNA: mining and meaning. *Curr. Opin. Struct. Biol.*, **8**, 333–337.
31. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
32. van Helden,J., Rios,A.F. and Collado-Vides,J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, **28**, 1808–1818.