Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W571–W577***
doi:10.1093/nar/gkl279

CRSD: a comprehensive web server for composite regulatory signature discovery

Chun-Chi Liu[1,2], Chin-Chung Lin[2], Wen-Shyen E. Chen[1], Hsuan-Yu Chen[3], Pei-Chun Chang[4], Jeremy J.W. Chen[2,5,*] and Pan-Chyr Yang[5]

[1]Department of Computer Science, National Chung-Hsing University, Taichung, Taiwan, ROC, [2]Institutes of Biomedical Sciences and Molecular Biology, National Chung-Hsing University, Taichung, Taiwan, ROC, [3]Graduate Institute of Epidemiology, National Taiwan University, Taipei, Taiwan, ROC, [4]Departments of Biotechnology and Bioinformatics, Asia University, Taichung, Taiwan, ROC, and [5]NTU Center for Genomic Medicine, National Taiwan University College of Medicine, Taipei, Taiwan, ROC

Received February 14, 2006; Revised March 19, 2006; Accepted April 4, 2006

ABSTRACT

Transcription factors (TFs) and microRNAs play important roles in the regulation of human gene expression, and the study of their combinatory regulations of gene expression is a new research field. We constructed a comprehensive web server, the composite regulatory signature database (CRSD), that can be applied in investigating complex regulatory behaviors involving gene expression signatures (GESs), microRNA regulatory signatures (MRSs) and TF regulatory signatures (TRSs). Six well-known and large-scale databases, including the human UniGene, mature microRNAs, putative promoter, TRANSFAC, pathway and Gene Ontology (GO) databases, were integrated to provide the comprehensive analysis in CRSD. Two new genome-wide databases, of MRSs and TRSs, were also constructed and further integrated into CRSD. To accomplish the microarray data analysis at one go, several methods, including microarray data pretreatment, statistical and clustering analysis, iterative enrichment analysis and motif discovery, were closely integrated in the web server, which has not been the case in previous studies. Our implementation showed that the published literature could demonstrate the results of genome-wide enrichment analysis. We conclude that CRSD is a powerful and useful bioinformatic web server and may provide new insights into gene regulation networks. CRSD and the online tutorial are publicly available at http://biochip.nchu.edu.tw/crsd1/.

INTRODUCTION

Transcriptional regulation of gene expression is mediated by binding of transcription factors (TFs) to specific regulatory DNA elements, typically upstream from and close to the transcription start point of a gene (1). Over the past few years, microRNAs have been shown to play a key role in gene regulation. Therefore, the combinatorial mechanisms of switch regulations composed by TFs and microRNAs make gene regulation networks more complex. Those complex regulatory networks include the same gene targeted by both a TF and microRNA (2). Most TFs have to recognize specific sequences within promoter regions to work, and these specific genomic sequences are known as TF binding sites. In contrast, microRNAs play negatively regulating roles in silencing gene functions in biomass (3). A mature microRNA with RISC complex will target specific mRNA with homology sequences complementary to the mature microRNA within the 3′-untranslated region (3′-UTR) (4).

A previous study predicted that there are 2273 genes with microRNA target sites conserved in mammals by scanning 3′-UTRs from the human, mouse and rat genomes for potential target sites (5). Moreover, in a five-genome (human, mouse, rat, chicken and dog) analysis of 3′-UTRs, ∼13 000 regulatory relationships were detected, which suggests that the conserved microRNAs might target more than 5300 human genes (6). However, some studies indicate that it is possible to reliably predict microRNAs without using genome comparisons (7,8). Because of a lack of detailed information on microRNA evolution, the conservation of target sites provides practical information on predicted target sites, and the new microRNAs have continuously appeared in possible evolution (5). Thus, we developed the prediction method that uses the human genome without relying on conservation. In order to perform genome-wide microRNA target prediction, the human

*To whom correspondence should be addressed. Tel: 886 4 22840485, ext. 226; Fax: 886 4 22853469; Email: jwchen@dragon.nchu.edu.tw

UniGene database (9) was employed to obtain the putative 3′-UTR database.

A previous report defined the transcriptional regulatory signature as the interactions between a TF and a group of genes with putative binding sites in the promoter sequences (10). We therefore defined the microRNA regulatory signature (MRS), the TF regulatory signature (TRS) and the composite regulatory signature (CRS) as follows: the MRS consists of the interactions of a microRNA and a group of genes with the putative targets of the former in 3′-UTR; the TRS is defined as the equivalent to the transcriptional regulatory signature, and the CRS is the combination of several MRSs and TRSs for a common group of genes. The MRS and TRS databases were integrated to establish a composite regulatory signature database (CRSD) that is also a comprehensive web server for CRS discovery.

DNA microarrays are often used to generate gene expression signatures (GESs) of tissues or cells under physiological or pathological conditions (11–13). GESs may include co-regulated groups of genes. Advanced enrichment analysis and motif discovery can identify these co-regulated groups. A recent report investigated cancer GESs for the enrichment of particular gene annotations and metabolic and signaling pathways (14) using Gene Ontology (GO) (15), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (16) and Biocarta pathways databases (http://www.biocarta.com/). In addition, each GES was also assessed for each significant enrichment of TRSs (10,17). Previous reports have combined the GESs, putative promoter sequences and GO annotations to investigate TF regulatory behavior (17,18), as well as to discover regulatory motif sequences (18). Enrichment analysis of MRSs is important but has not been performed. The assessment of all MRSs and TRSs for the significant enrichment of all GO annotations, KEGG pathways and Biocarta pathways has not been carried out either.

In this study, these important enrichment analyses were performed, and an iterative enrichment analysis model and user-friendly web interface were developed, providing novel methods to yield comprehensive information in the field of such complex large-scale databases. We also provided a user-friendly interface for researchers to perform queries in CRSD easily, improving system performance using a hash table algorithm and cache technology. CRSD is a novel and comprehensive web server that closely integrates several methods, including microarray data pretreatment, microarray data statistical and clustering analysis, genome-wide iterative enrichment analysis and motif discovery, which has not been the case in previous studies.

MATERIALS AND METHODS

Database construction

The MRS database was constructed using mature human microRNAs (19) and 3′-UTR sequences, and the TRS database was constructed using TRANSFAC (20) and the promoter database PromoSer (21). The detailed procedures are described in Supplementary Data.

The framework of the CRSD web server

The CRSD has four major functional components: (i) microarray data pretreatment, (ii) microarray data statistical and clustering analysis, (iii) iterative enrichment analysis and (iv) motif discovery. Figure 1A shows the high-level of the four major functional components, and Figure 1B shows the detailed workflows. Users can obtain a preliminary result in the microarray data pretreatment component and later perform further analysis such as microarray data statistical and clustering analysis, enrichment analysis and motif discovery (Figure 1A). The CRSD and the online tutorial are publicly available at http://biochip.nchu.edu.tw/crsd1/.

Microarray data analysis and GES

Microarray data analysis has two major components: microarray data pretreatment and microarray data statistical/clustering analysis. The former includes quantile normalization (rescaling) (22), data adjustment, data filtration and standard normalization; the latter includes student's t-test, signal-to-noise test, ANOVA test and self-organizing map (SOM) clustering (23). Student's *t*-test and the signal-to-noise test can determine the group-specific marker genes that can be treated as a GES. The group-specific marker genes correlating with one particular group versus all other groups were identified using the *P*-value according to the statistical test. The permutation test was mainly applied to the marker gene selection using the signal-to-noise approach; however, it also can be used with student's *t*-test. To adjust the *P*-value for a multiple hypothesis test, the false discovery rate (*Q*-value) was estimated using the method in a previous study (10). An example of microarray data pretreatment and analysis, including quantile normalization, data adjustment, data filtration, standard normalization and student's *t*-test, is given in Supplementary Figure S1.

Enrichment analysis

CRSD provides various types of enrichment analyses. Supplementary Figure S2 shows an instance of enrichment analysis for one group of genes (G_A) for the significant enrichment of another (G_B). G_A or G_B could be the genes belonging to a GO annotation, pathway, MRS, TRS, GES or user input data. The possible set P_A is defined as the set of all possible genes of G_A. We counted the number of genes intersecting G_A and G_B: $n = c(G_A \cap G_B)$, where $c(X)$ denotes the number of elements in set X. If G_A is a GES, P_A will be the set of all detectable genes in the microarray. If G_A is a TRS, P_A will be 23 095 genes with promoter sequences. If G_A is an MRS, P_A will be 54 576 genes with a 3′-UTR. If users do not have the exact information about the possible set *of* G_A, they can consider assigning the 54 576 genes of UniGene to P_A. Next, we calculated the probability of observing an equal or larger intersection between G_A and G_B by chance by summing the binomial distribution probabilities (10) for all intersections of equal or larger size:

$$P\text{-value} = \sum_{i=n}^{b} \left(\frac{b!}{(b-i)!i!} \right) \left(\frac{a}{N} \right)^{i} \left(1 - \frac{a}{N} \right)^{b-i} \quad \text{where } N = c(P_A), a = c(G_A), b = c(G_B) \qquad \mathbf{1}$$

If the enrichment analyses are to assess G_A for $G_{B1}, G_{B2}, \ldots, G_{BK}$, the *P*-value is calculated for each G_B. Then, in order to adjust the *P*-value for the multiple hypothesis test, the

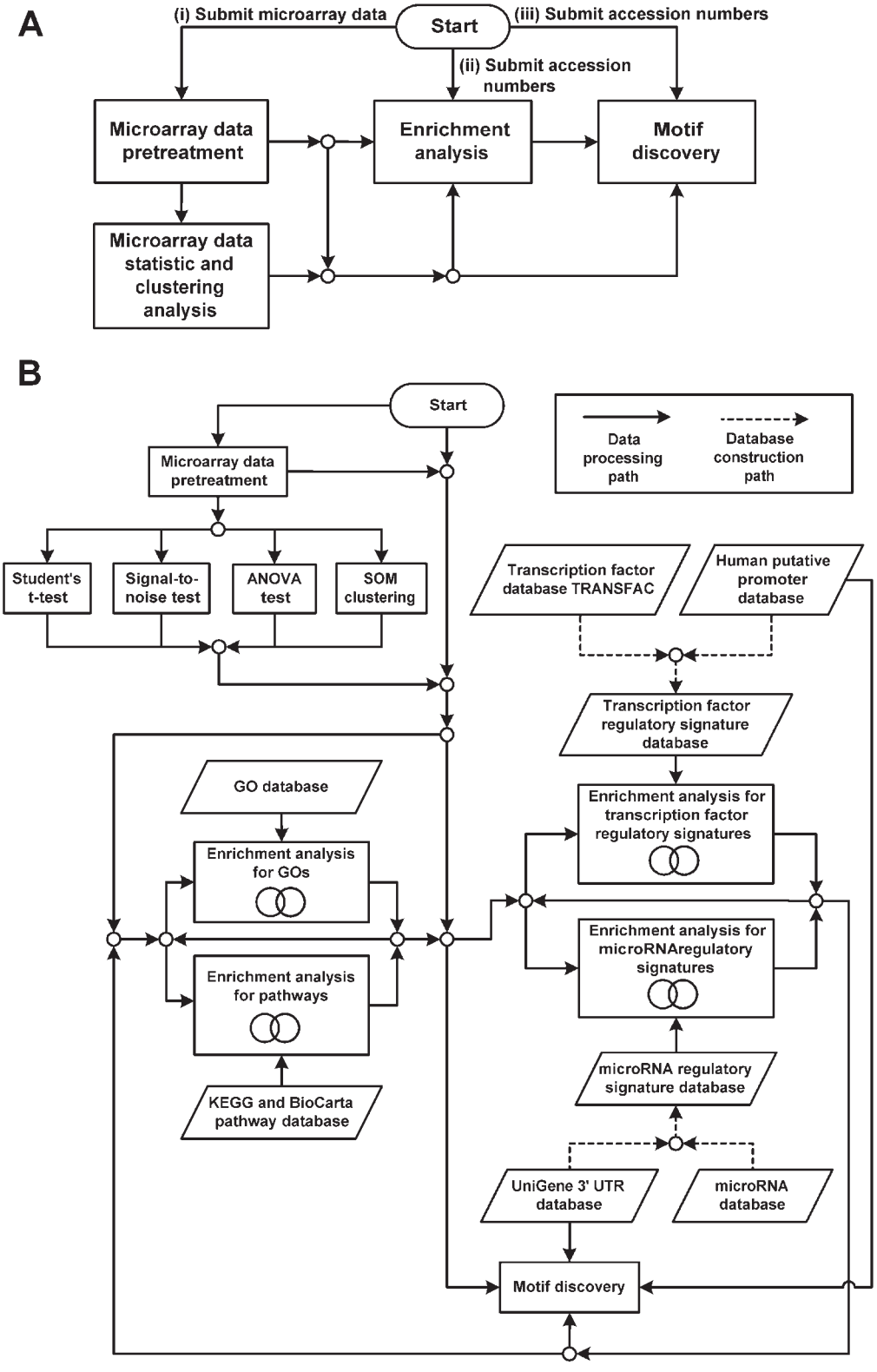

Figure 1. The architecture of CRSD. (**A**) The high-level workflow for the four main functional components: (i) microarray data pretreatment, (ii) microarray data statistical and clustering analysis, (iii) iterative enrichment analysis and (iv) motif discovery. CRSD provides three initial procedures using microarray data pretreatment, enrichment analysis and motif discovery. (**B**) The MRS database integrating the microRNA and 3′-UTR databases is constructed by microRNA target prediction, and the TRS database is constructed by TF binding site prediction integrating promoter and TRANSFAC databases. The detailed data processing paths are shown in the flowchart, which represents the iterative enrichment analysis among GO, pathway, TRS and MRS.

Q-value (false discovery rate) was calculated using the method described previously (10), as follows:

$$Q\text{-value} = \frac{K \times P\text{-value}}{R} \quad \mathbf{2}$$

Where K is the number of G_B tested against G_A, and R is the ascending order rank of the respective P-value. The alignment score of the putative microRNA target and the similarity score of the TF binding site were not considered in enrichment analysis. Therefore, we defined the target score (T-Score) to evaluate the significance of the intersection targets, which is described in the Supplementary Data.

Novel motif discovery

To identify the highly conserved sequences for the novel TF binding sites in the putative promoter sequences of a group of input genes, the promoter database and motif discovery tools, including Gibbs sampling-based GLAM (24) and information theory-based Weeder (25), were employed. The same strategy was used to predict the novel targeting sites of microRNAs, conserved at the 3′-UTR of a group of genes, using our 3′-UTR database and the motif discovery tools. The motif discovery component was closely integrated with the enrichment analysis and microarray data analysis, providing a user-friendly interface for CRSD to efficiently utilize these comprehensive components.

RESULTS

Database construction

Following these database construction methods, 54 576 putative 3′-UTRs, with an average length of 744 bases, were obtained. These sequences were used to create the putative 3′-UTR database. After the application of threshold and rank filtration procedures, we constructed the MRS database which contained 319 MRSs and 615 267 potential regulatory interactions, and the TRS database which contained 494 TRSs and 928 592 potential regulatory interactions.

Design and features of the CRSD

CRSD utilized and integrated six well-known, large-scale databases, including the human UniGene, mature microRNAs, putative promoter, TRANSFAC, pathway and GO databases, to provide a comprehensive knowledge-base. In addition, two new genome-wide databases, the MRS and TRS databases were constructed. Users can submit data to CRSD using three workflows: (i) submitting microarray data to the microarray data pretreatment component, (ii) submitting the GenBank accession numbers or UniGene IDs of a group of genes to the enrichment analysis component and (iii) submitting the GenBank accession numbers or UniGene IDs of a group of genes to the motif discovery component. Figure 1A shows these three initial workflows. CRSD also provides a framework that integrates microarray data analysis, enrichment analysis and motif discovery (Figure 1B). Examples of server usage are described in the Supplementary Data.

Iterative enrichment analysis

There are five types of enrichment analysis in CRSD: CRSs, MRSs, TRSs, pathways and GO annotations. For example, a group of genes can be the input to the enrichment analysis components for CRS discovery. Then, genome-wide enrichment analysis for all MRSs and TRSs is performed, and the significant enrichment MRSs and TRSs are output. These output microRNAs and TFs have statistically significant enrichment with the input genes. It is possible that these microRNAs and TFs can perform gene regulation via a combinatorial mechanism or act cooperatively on their targets (the input genes).

Partial screenshots of the results page for the enrichment analysis for MRSs, TRSs, pathways and GO annotations are shown in Figure 2A–D. Users can continually investigate the intersection between the input genes and a significant enrichment group that could be an MRS, TRS, pathway or GO annotation. Thus, we provide iterative enrichment analysis and motif discovery for the selected genes of each intersection in the previous enrichment analysis. For example, carrying out the discovery of the significant enrichment MRS for a group of genes (Figure 2A), users can perform other functional options, including GO annotation, pathway, TRS and motif discovery.

DISCUSSION

DNA microarrays are a powerful tool for massive parallel analysis of gene expression and have been applied in various biological studies in the post-genomic era to identify GESs (11–13). How to interpret and characterize these expression signatures is always perplexing to researchers. In the past few years, it has become increasingly clear that gene regulation can be modulated not only by TFs but also by microRNAs (2). To resolve the complex regulatory behaviors of gene expression, the development of a comprehensive and powerful bioinformatics tool is absolutely necessary. Therefore, we developed CRSD which can be applied to investigate complex regulatory behaviors involving GESs, MRSs and TRSs. In order to discover composite regulatory behaviors, CRSD provides several functions. (i) Microarray data analysis can be utilized to identify the GESs. (ii) Genome-wide iterative enrichment analysis can be utilized to discover the significant enrichment microRNAs and TFs associated with the GESs. The predicted TFs and microRNAs can be further confirmed by electrophoretic mobility shift assays (and/or reporter assays) and real-time RT–PCR, respectively. (iii) Motif discovery can be employed to predict the TF binding sites and microRNA targeting sites for each group of genes in the intersection between the GES and one of the significant MRSs, TRSs, pathways or GO annotations. The predicted TF binding sites and microRNA targeting sites can be used to discover novel TFs and microRNAs. There are many scenarios in this workflow. For example, if a user identifies a GES for lung cancer tissues and finds a significant enrichment pathway associated with this GES, then the novel TF binding sites of the genes in the intersection between the GES and the pathway can be discovered using the motif discovery function in CRSD. The results may imply that the target genes with the predicted TF binding site are co-regulated in lung cancer and related to the predicted pathway.

A. Significant MicroRNA Regulatory Signatures

Rank	microRNA	Total	Found	P-Value	Q-Value	microRNA Targets	Enrichment and Motif
1	hsa-miR-500	1940	4	6.281E-4	8.479E-2	Hs.512867 (H63) score 104 energy -27.14 Query: 3' GUCU--UAGG--AACGGGUCCACGUA 5' Ref: 5' CAGGTTGTTCACTTGCTGAGGTGCAA 3' Hs.125300 (TRIM34) score 111 energy -21.62 Query: 3' GUCUUAGGAACGGGUCCACGUA 5' Ref: 5' CAGGATAC---CCCAGGTACAT 3' Hs.130635 () score 107 energy -21.89 Query: 3' GUCUUAGGAACGGGUCCACGUA 5' Ref: 5' CAGAACTTTTGTTCAGATGTAA 3' Hs.105468 (GIMAP6) score 104 energy -22.37 Query: 3' GUCUUAGGA-AC-GGGUCCACGUA 5' Ref: 5' GAGACTGATGTGCCCCGGGTGCAC 3'	GO Pathway TF Motif

B. Significant Transcription Factor Regulatory Signatures

Rank	Factor Name (Matrix) Description	Total	Found	P-Value	Q-Value	Transcription Factor Binding Sites (UniGeneID Position Similarity Sequence)	Enrichment and Motif
1	CHX10 (V$CHX10_01)	1773	8	1.329E-5	5.728E-3	Hs.311776 (TCEAL3) -1656 (-) 0.944 ctgtctAATTAtcc Hs.165830 (EVI1) -134 (+) 0.994 aaaTAATTagcttt Hs.475629 (TBC1D5) -895 (-) 0.944 cactctAATTAaca Hs.435342 (SLU7) -701 (-) 0.997 ttggctAATTAggc Hs.445497 (HNRPLL) -1362 (-) 0.996 aatgctAATTAtct Hs.4055 (KLF6) -1491 (-) 0.995 agcgctAATTActg Hs.209431 (MGC3794) -1053 (-) 0.997 acagctAATTAaca Hs.151411 (MYCBP2) -1308 (-) 0.996 actgctAATTAtct Hs.151411 (MYCBP2) -1375 (-) 0.994 agggctAATTAcag	GO Pathway microRNA Motif

C. Significant Pathways

Rank	ID	Type	Name	Total	Found	P-Value	Q-Value	Associated Genes (UniGene)	Enrichment and Motif
1	h_alk	BioCarta	ALK in cardiac myocytes	36	4	1.798E-4	8.775E-2	Hs.476018 (CTNNB1) Hs.445733 (GSK3B) Hs.133379 (TGFB2) Hs.494622 (TGFBR1)	TF microRNA GO Motif

D. Significant GOs

Rank	ID	Type	Name	Total	Found	P-Value	Q-Value	Associated Genes (UniGene)	Enrichment and Motif
1	GO:0015031	GO	protein transport	214	9	2.411E-6	8.799E-4	Hs.513057 (RANBP5) Hs.462742 (RHOT1) Hs.432755 (SNAG1) Hs.476930 (DKFZP564O123) Hs.467824 (PUM2) Hs.523470 (IPO7) Hs.549125 (RHOQ) Hs.524574 (NUP107) Hs.532793 (KPNB1)	TF microRNA Pathway Motif

Figure 2. Partial screenshots of the results page for the enrichment analysis. (**A**) The results page for significant enrichment MRSs showing the total number of target genes of microRNA hsa-miR-500, the retrieved set (intersection) of input genes, *P*-value, *Q*-value, microRNA targets and additional analysis buttons, including GO annotation, pathway, TRS and motif discovery of the interaction genes. (**B**) The results page for significant enrichment TRSs showing the total number of target genes, the retrieved set of input genes, *P*-value, *Q*-value, TF binding sites, and additional analysis buttons. (**C**) The results page for significant enrichment pathways showing the total number of pathway genes, the retrieved set of input genes, *P*-value, *Q*-value, the interaction genes, and additional analysis buttons. (**D**) The results page for significant enrichment GO annotations showing the total number of genes of the GO annotation, the retrieved set of input genes, *P*-value, *Q*-value, the interaction genes and additional analysis buttons.

In this study, genome-wide enrichment analysis was performed for the CRS prediction of pathways and GO annotations. A total of 159 human KEGG pathways and 293 human Biocarta pathways were processed, and we report the significant MRSs and TRSs for each pathway in Supplementary Table S1. In addition, a total of 4479 GO annotations related to the human genome were processed, and we report the significant MRSs and TRSs for each GO annotation in Supplementary Table S2. The *T*-Score was calculated in these analyses and is reported in both tables. Some previous studies

have demonstrated the predictions of microRNA targets performed in CRSD. For instance, microRNA hsa-miR-15a and hsa-miR-7 can regulate *BCL2* (26) and *EGFR* (27), respectively. Our MRS database also predicted that hsa-miR-15a and hsa-miR-7 can regulate *BCL2* and *EGFR*, respectively. Previous studies also have supported our predicted TFs having significant enrichment with the pathway (Supplementary Table S3); the results for predicted TFs are in detail in the Supplementary Data.

The TRANSFAC database is the most commonly used repository for TF binding sites (28) and has been used in many studies (1,10,18). We used the matrices for the vertebrate group in the TRANSFAC database that are cross-species motifs. The identification of species-specific motifs may be an important issue in TF binding site discovery; however, we cannot currently find an accurate and suitable human-specific motif database. Further analyses of species-specific motifs may be necessary.

In this study, each promoter in the human putative promoter database contains 2000 bases upstream of and 100 bases downstream of the transcription start point. However, previous studies have demonstrated that TF binding sites might be located downstream of transcription start points (29,30) beyond the region we have reported; indeed, the current version of server cannot cover the TF binding sites in the intron, 3′-UTR, or a region distant from the transcription start point, which is a limitation of this web server. Other regions such as 3′-UTRs and introns for TF binding sites should be taken into account in future work.

In summary, we have developed a user-friendly interface and powerful platform for researchers to carry out their work easily. CRSD closely integrates the approaches of microarray data analysis, genome-wide iterative enrichment analysis and motif discovery with both well-known public databases and ours. By means of CRSD analyses, investigators can explore the complex regulatory behaviors involving GESs, MRSs and TRSs and may obtain new insights into gene regulation networks.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by the National Chung-Hsing University Biotechnology Center through the Teaching Core Facility Project grant. Funding to pay the Open Access publication charges for this article was provided by National Chung-Hsing University.

Conflict of interest statement. None declared.

REFERENCES

1. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
2. O'Donnell,K.A., Wentzel,E.A., Zeller,K.I., Dang,C.V. and Mendell,J.T. (2005) c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*, **435**, 839–843.
3. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
4. Lin,S.Y., Johnson,S.M., Abraham,M., Vella,M.C., Pasquinelli,A., Gamberi,C., Gottlieb,E. and Slack,F.J. (2003) The *C. elegans* hunchback homolog, hbl-1, controls temporal patterning and is a probable microRNA target. *Dev. Cell*, **4**, 639–650.
5. John,B., Enright,A.J., Aravin,A., Tuschl,T., Sander,C. and Marks,D.S. (2004) Human MicroRNA targets. *PLoS Biol.*, **2**, e363.
6. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
7. Lindow,M. and Krogh,A. (2005) Computational evidence for hundreds of non-conserved plant microRNAs. *BMC Genomics*, **6**, 119.
8. Bentwich,I., Avniel,A., Karov,Y., Aharonov,R., Gilad,S., Barad,O., Barzilai,A., Einat,P., Einav,U., Meiri,E. *et al.* (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genet.*, **37**, 766–770.
9. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
10. Rhodes,D.R., Kalyana-Sundaram,S., Mahavisno,V., Barrette,T.R., Ghosh,D. and Chinnaiyan,A.M. (2005) Mining for regulatory programs in the cancer transcriptome. *Nature Genet.*, **37**, 579–583.
11. Chen,J.J., Lin,Y.C., Yao,P.L., Yuan,A., Chen,H.Y., Shun,C.T., Tsai,M.F., Chen,C.H. and Yang,P.C. (2005) Tumor-associated macrophages: the double-edged sword in cancer progression. *J. Clin. Oncol.*, **23**, 953–964.
12. Chen,J.J., Peck,K., Hong,T.M., Yang,S.C., Sher,Y.P., Shih,J.Y., Wu,R., Cheng,J.L., Roffler,S.R., Wu,C.W. *et al.* (2001) Global analysis of gene expression in invasion by a lung cancer model. *Cancer Res.*, **61**, 5223–5230.
13. Wang,C.C., Tsai,M.F., Hong,T.M., Chang,G.C., Chen,C.Y., Yang,W.M., Chen,J.J. and Yang,P.C. (2005) The transcriptional factor YY1 upregulates the novel invasion suppressor HLJ1 expression and inhibits cancer cell invasion. *Oncogene*, **24**, 4081–4093.
14. Rhodes,D.R. and Chinnaiyan,A.M. (2005) Integrative analysis of the cancer transcriptome. *Nature Genet.*, **37**, S31–S37.
15. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
16. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
17. Haverty,P.M., Hansen,U. and Weng,Z. (2004) Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic Acids Res.*, **32**, 179–188.
18. Yap,Y.L., Lam,D.C., Luc,G., Zhang,X.W., Hernandez,D., Gras,R., Wang,E., Chiu,S.W., Chung,L.P., Lam,W.K. *et al.* (2005) Conserved transcription factor binding sites of cancer markers derived from primary lung adenocarcinoma microarrays. *Nucleic Acids Res.*, **33**, 409–421.
19. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
20. 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.
21. Halees,A.S. and Weng,Z. (2004) PromoSer: improvements to the algorithm, visualization and accessibility. *Nucleic Acids Res.*, **32**, W191–W194.
22. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
23. Davide,F.A., Di Natale,C. and D'Amico,A. (1995) Self-organising sensory maps in odour classification mimicking. *Biosens. Bioelectron.*, **10**, 203–218.

24. Frith,M.C., Hansen,U., Spouge,J.L. and Weng,Z. (2004) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res*., **32**, 189–200.
25. Pavesi,G., Mereghetti,P., Mauri,G. and Pesole,G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res*., **32**, W199–W203.
26. Cimmino,A., Calin,G.A., Fabbri,M., Iorio,M.V., Ferracin,M., Shimizu,M., Wojcik,S.E., Aqeilan,R.I., Zupo,S., Dono,M. *et al*. (2005) miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc. Natl Acad. Sci. USA*, **102**, 13944–13949.
27. Li,X. and Carthew,R.W. (2005) A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the *Drosophila* eye. *Cell*, **123**, 1267–1277.
28. Fogel,G.B., Weekes,D.G., Varga,G., Dow,E.R., Craven,A.M., Harlow,H.B., Su,E.W., Onyia,J.E. and Su,C. (2005) A statistical analysis of the TRANSFAC database. *Biosystems*, **81**, 137–154.
29. Gaubatz,S., Meichle,A. and Eilers,M. (1994) An E-box element localized in the first intron mediates regulation of the prothymosin alpha gene by c-myc. *Mol. Cell. Biol*., **14**, 3853–3862.
30. Beohar,N. and Kawamoto,S. (1998) Transcriptional regulation of the human nonmuscle myosin II heavy chain-A gene. Identification of three clustered cis-elements in intron-1 which modulate transcription in a cell type- and differentiation state-dependent manner. *J. Biol. Chem*., **273**, 9168–9178.