Genomics & Informatics Vol. 5(1) 10-18, March 2007

FCAnalyzer: A Functional Clustering Analysis Tool for Predicted Transcription Regulatory Elements and Gene Ontology Terms

Sang-Bae Kim[1†], Gil-Mi Ryu[2†], Young-Jin Kim[2†], Jee-Yeon Heo[2], Chan Park[2], Bermseok Oh[2], Hyung-Lae Kim[2], Kuchan Kimm[2], Kyu-Won Kim[3*] and Young-Youl Kim[2*]

[1]Korean BioInformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejon 305-806, [2]Center for Genome Science, National Institute of Health, Seoul 122-701, Korea, [3]College of Pharmacy, Seoul National University, Seoul 157-742, Korea

Abstract

Numerous studies have reported that genes with similar expression patterns are co-regulated. From gene expression data, we have assumed that genes having similar expression pattern would share similar transcription factor binding sites (TFBSs). These function as the binding regions for transcription factors (TFs) and thereby regulate gene expression. In this context, various analysis tools have been developed. However, they have shortcomings in the combined analysis of expression patterns and significant TFBSs and in the functional analysis of target genes of significantly overrepresented putative regulators. In this study, we present a web-based A Functional Clustering Analysis Tool for Predicted Transcription Regulatory Elements and Gene Ontology Terms (FCAnalyzer). This system integrates microarray clustering data with similar expression patterns, and TFBS data in each cluster. FCAnalyzer is designed to perform two independent clustering procedures. The first process clusters gene expression profiles using the K-means clustering method, and the second process clusters predicted TFBSs in the upstream region of previously clustered genes using the hierarchical biclustering method for simultaneous grouping of genes and samples. This system offers retrieved information for predicted TFBSs in each cluster using Match™ in the TRANSFAC database. We used gene ontology term analysis for functional annotation of genes in the same cluster. We also provide the user with a combinatorial TFBS analysis of TFBS pairs. The enrichment of TFBS analysis and GO term analysis is statistically by the calculation of *P* values based on Fisher's exact test, hypergeometric distribution and Bonferroni correction. FCAnalyzer is a web-based, user-friendly functional clustering analysis system that facilitates the transcriptional regulatory analysis of co-expressed genes. This system presents the analyses of clustered genes, significant TFBSs, significantly enriched TFBS combinations, their target genes and TFBS-TF pairs.

Availability: This tool is freely available for academic and nonprofit users at http://www.ngri.go.kr/cgi-bin/cmams/fcanalyzer.cgi

Keywords: Transcription regulatory element, clustering analysis

Introduction

Microarray technology is a powerful analytical tool that is commonly used to elucidate the expression patterns of coordinately regulated genes. Microarrays have provided a revolutionary platform for the study of gene expression, regulation and function (Walsh and Henderson, 2004). They have tremendous potential for the study of biological processes associated with health and disease. Microarray-based gene expression analysis will undoubtedly make a major contribution to our understanding of the underlying biological mechanisms of disease, and will ultimately lead to improved methods for the diagnosis, prognosis and treatment of disease (Sausville and Holbeck, 2004). The expression data produced by microarray hybridization experiments can lead to the identification of clusters of co-expressed genes that are likely to be co-regulated by shared regulatory mechanisms. Many types of high-throughput functional genomic data that can facilitate rapid functional annotation of sequenced genomes are currently available. Due to the considerable quantity and intrinsic variation of the data produced from microarray experiments, statistical and computational approaches, including clustering analysis, have been used to consolidate useful biological information from microarray data (Lobenhofer *et al.*, 2001).

Moreover, the wide availability of genomic data, including sequences in the upstream regions of genes, has

†These authors contributed equally.
*Corresponding authors: E-mail youngyk@nih.go.kr, gwonkim@plaza.snu.ac.kr
Tel +82-2-380-2245, Fax +82-2-358-1063
Accepted 19 Dec 2006

made it possible to perform large-scale studies combining gene expression data, and to enable researchers to understand both the regulation and the mechanism of gene transcription. By analyzing the upstream promoter regions of coexpressed genes, it is possible to elucidate common regulatory patterns characterized by TFBSs (Kellis *et al.*, 2003). However, it is important to realize that precise analysis methods are necessary to make an accurate functional interpretation of these large-scale data sets.

The regulation of transcription is a very complex process in higher eukaryotic organisms. A large number of nuclear TFs control gene expression by binding to regulatory sequence elements that are located upstream from the transcriptional start sites of genes (Villard, 2004). TFBSs are usually 5-25 base pairs in length, and are often represented as matrices. These matrices are referred to in the literature under a variety of labels such as position weight matrices (PWM), position frequency matrices (PFM), alignment matrices, profiles and so on (Knuppel *et al.*, 1994), (Sandelin *et al.*, 2004). Many TFBS prediction programs using PWM, such as promoterscan (Prestridge, 1995), TSSG, and TSSW (Solovyev and Salamov, 1997) have been developed (Murakami *et al.*, 2004).

In order to explain the mechanisms of gene expression in detail, it is crucial that we understand the effects of TFBSs on transcriptional regulation at a genomic level, as well as gene expression patterns at a transcriptional level (Kasturi and Acharya, 2005). Analysis tools which function on a genomic level have previously been developed (Roth *et al.*, 1998; Sinha and Tompa 2003; Liu et al., 2004; Kim *et al.*, 2005); however, they have limitations as they are unable to perform this kind of combined data analysis. Recently, the analysis tool EXPANDER was introduced, which is a java package for the analysis of gene expression data, and provides multiple-functions such as clustering, visualization, biclustering and downstream analysis such as functional enrichment and promoter analysis (Shamir *et al.*, 2005). Although it offers a variety of functional analyses for genes, it does not provide perspective functional analysis between co-expressed genes, their significant TFBSs and the function of target genes. FCAnalyzer focuses more on verifying the relationship between the regulatory region and gene expression. Additionally, we verified enrichment for TFBSs, TFBS combination and functional annotations in a cluster based on *P* values.

The aim of this study was to offer a novel tool which operates to perform both clustering analysis for the detection of expression patterns and validation of the transcriptional regulatory element patterns from coexpressed genes. We introduce a web-based functional clustering analysis tool that integrates microarray data with similar expression patterns and TFBS data to identify novel transcription regulatory networks of known genes in the human, mouse and rat genomes. FCAnalyzer performs two-step clustering procedures; K-means and hierarchical clustering. K-means clustering is one of the simplest unsupervised learning algorithms used to classify, or group objects together based on attributes/features, into K number of groups, where K is a positive integer. The grouping is done by minimizing the sum of the squares of distances between data and the corresponding cluster centroid. Thus the purpose of K-means clustering is to classify the data (Tavazoie *et al.*, 1999). The first procedure clusters gene expression profiles using the K-means clustering method, and then applies the hierarchical method (Eisen *et al.*, 1998) for predicted TFBSs clustering of genes that were previously clustered. The hierarchical method is one of the most popular analytical methods currently being used to characterize gene-expression. The combined clustering approach that we have developed enables researchers to obtain results with more biological significance than those obtained by analysis with either TFBSs or gene expression data alone. Using TFBS analysis with the clusters that exhibited similar expression patterns, we obtained TF binding and TFBS information. This information was used to construct a matrix representing TFBS-TF interactions. We expect this tool to be useful in understanding transcription mechanisms and in the characterization of genetic networks.

In the cluster analysis of gene expression profiles, the functional annotation of the cluster is the most prominent method for identifying functions of co-expressed genes. In order to accomplish the functional annotation of genes within each cluster, we utilized an efficient computational approach that relies on the description of biological processes, molecular functions or cellular components using GO terms (Ashburner *et al.*, 2000).

Materials and Methods

We developed FCAnalyzer which functions as a web-based clustering analysis tool with predicted TFBSs of genes which show similar expression patterns generated from microarray data. FCAnalyzer is implemented in R language (http://www.bioconductor.org) and uses the MySQL (http://www.mysql.com) database. This program is enveloped in a Perl script (http://www.perl.com) in order to maintain a user friendly web interface. FCAnalyzer is accessible at http://www.ngri.go.kr/cgi-bin/cmams/fcanalyzer.cgi.

FCAnalyzer consists of three sections:

Databases

Construction of TFBS matrix databases for known genes of the human, mouse and rat

The TFBS matrix database consists of putative promoter regions of upstream DNA sequences around 2,000 bp from the transcription start site. In order to construct the TFBS matrix database, we first collected promoter sequences of known Refseq genes. The genome assemblies used for promoter sequence resources were Build 35 in human, Build 33 in mouse and Rnor3.4 in rat. These data were obtained from the Genome sequencing consortium (http://www.hgsc.bcm.tmc.edu/). We obtained the 2,000 bp promoter region from the UCSC Genome Browser (http://genome.ucsc.edu/) and used the Match™ matrix based search program to predict TFBSs. Match™ is a powerful web-based tool which uses positional weight matrices in TRANSFAC® Professional. To obtain high fidelity TFBS matrices, we set the parameters of Match™ to minimize false positives. After analyzing TFBS frequencies for each gene, we generated a TFBS frequency matrix in a species. Rows of the matrix are defined as Refseq genes while the columns represent predicted TFBSs. We present here a novel matrix database containing the results following the application of the TFBS frequency matrix. The matrix identified 343 TFBS for 20233 genes in human, 360 TFBS for 13032 genes in mouse and 320 TFBS for 3009 in rat.

Construction of TF, TFBS and GO information databases

Databases were constructed to present TF and TFBS information in designated clusters. The annotation was based on the annotation used in the TRANSFAC® database, version 8.4. To facilitate the functional annotation of cluster analysis, we constructed a GO annotation database for Refseq genes. The GO project began with the overall goal of providing a unified view of gene functional annotations for different organisms with a set of three structured vocabularies; biological process, molecular function and cellular component. The structure of the ontology is based on directed acyclic graphs (DAGs) and organized in a hierarchy. GO information is useful for the functional analysis of specific gene sets such as clustered genes with same expression patterns (Martin *et al.*, 2004). We obtained all known Refseq information from NCBI (Jenuth 2000; Pruitt *et al.*, 2003). All of the ontology data and the gene-GO term associations were taken from the GO consortium website and the database for annotation, visualization and integrated discovery (DAVID) (Dennis *et al.*, 2003).

Clustering analysis

FCAnalyzer performs two-step clustering procedures: K-means clustering of gene expression data and the subsequent hierarchical clustering with predicted TFBS data for each cluster (Fig. 1).

Clustering of gene expression profiles

FCAnalyzer reads tab-delimited text files containing gene expression profiles with Refseq ID or Unigene ID in rows,

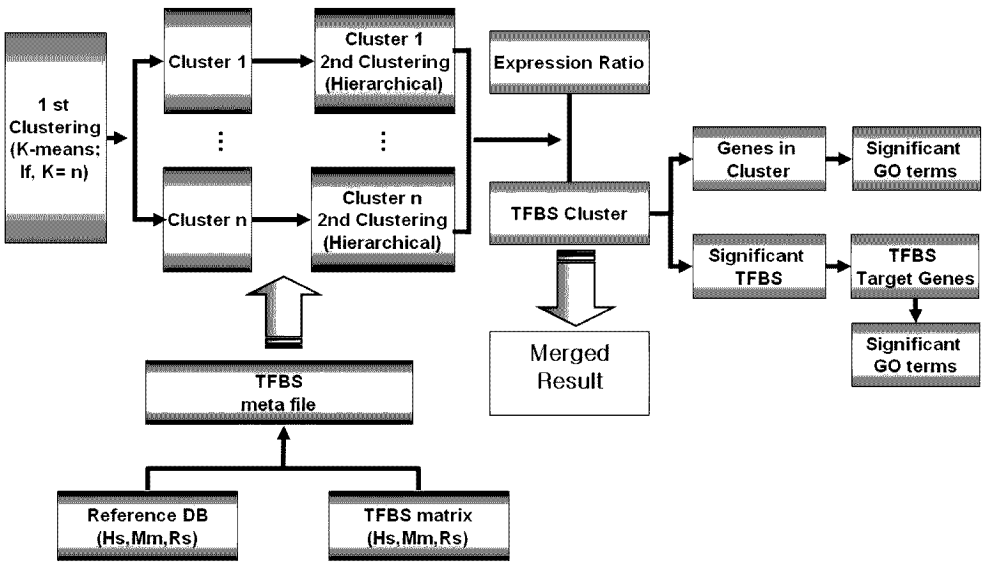


Fig. 1. Clustering analysis procedure. FCAnalyzer performs two-step clustering procedures following the functional analysis. The first step clusters a gene expression profile with the K-means clustering method and the second procedure clusters TFBSs in genes included in one of the previous clusters. Users can proceed to the functional analysis of significant TFBS and GO terms.

and experiments in columns. Users can set clustering parameter options for two-step clustering methods. The first setting is the K-means for gene expression profiles, and the second is the hierarchical setting for a predicted TFBS matrix. To aim at superior clustering results, users can apply other clustering methods (e.g. hierarchical clustering) in the preanalysis module in FCAnalyzer. The pre-analysis module provides various analyses of microarray data, from raw normalization, to higher analysis approaches such as significant gene selection, clustering, classification and visualization. Before the functional analysis, users can evaluate microarray data with the pre-process step of FCAnalyzer and the clustering result is stored in a database. In addition, a new TFBS matrix profile for clustered genes based predicted TFBS databases is created for each cluster.

Clustering of predicted TFBSs of genes in a cluster

The newly created TFBS profile includes all of the predicted TFBS information for each gene in a cluster. FCAnalyzer performs consecutive 2-way hierarchical clustering with the profile. This step is performed in order to find meaningful groups that share a similar TFBS pattern in genes that exhibit the same expression pattern. For the hierarchical clustering of the TFBS matrix, users can choose one of many different dissimilarity and distance matrix options to establish optimal clustering conditions. The linkage measures are Ward's minimum variance, complete, single, average, median, centroid and Mcquitty. The default distance measure is binary and others are also available such as euclidean, maximum, manhattan, canberra, uncentered pearson and centered pearson.

Following the two-step clustering analysis, FCAnalyzer produces heatmap figures showing TFBS clustering for genes in each cluster, and heatmaps for their respective gene expression profiles. In addition, it also provides tab-delimited text files for the result figures.

Exploring the biological relevance of clusters

Enrichment Analysis of TFBS

FCAnalyzer provides significant TFBS analysis and combinatorial TFBS analysis based on Fisher's exact test to identify statistically over-representative TFBSs located in the upstream sequences of genes in a cluster, and statistically significant enriched TFBS pairs in which two designated TFBSs are shown in the cluster. We used the test with a one-sided option, to compare the number of occurrences and nonoccurrences in genes of a cluster versus the all genes (Fig. 2 and 3). This process can help the user to understand how gene expression is regulated and how biological information is controlled at the transcriptional level.

TFBS-TF analysis

Analysis of target genes binding significant TFBSs

FCAnalyzer extracts functional annotation based on GO term analysis for potential target genes that are regulated by a specified set of TFs against significant TFBSs in a cluster. Functional categorization of the target genes may be very important in the deduction of the function of co-expressed genes, and in deciphering the transcriptional regulatory mechanisms in a cluster (Sosinsky *et al.*, 2003).

TFBS Analysis Result (ordered by P-value)

Download Result File

TFBS	No. genes in the cluster	No. genes in the hs genome	P-value (Fisher's test)
V$AMEF2_Q6	7/85	217/20233	0.00004
V$SMAD4_Q6	13/85	1032/20233	0.00034
V$MAZ_Q6	48/85	7765/20233	0.00051
V$PTF1BETA_Q6	4/85	149/20233	0.00355
V$BRN2_01	4/85	175/20233	0.00627
V$SRF_Q5_01	12/85	1301/20233	0.00779
V$POLY_C	2/85	40/20233	0.01217
V$RSRFC4_01	1/85	3/20233	0.01250
V$FOXD3_01	28/85	4480/20233	0.01368
V$MEF2_Q6_01	22/85	3290/20233	0.01503

Fig. 2. TFBS analysis result. The user can see the significant TFBS analysis result. The lists are sorted by p-values.

TFBS Analysis Result (ordered by P-value)

Download Result File

Combinatorial TFBS	P-value (Fisher's test)
V$AMEF2_Q6 V$DR4_Q2	0.00000
V$AMEF2_Q6 V$COMP1_01	0.00003
V$CHOP_01 V$LPOLYA_B	0.00005
V$COREBINDINGFACTOR_Q6 V$PTF1BETA_Q6	0.00009
V$AMEF2_Q6 V$VMAF_01	0.00009
V$DR4_Q2 V$MAZ_Q6	0.00010
V$AP4_01 V$NFKAPPAB50_01	0.00012
V$AMEF2_Q6 V$FOXD3_01	0.00012
V$FOXD3_01 V$SMAD4_Q6	0.00016
V$DR4_Q2 V$SREBP1_Q6	0.00020

Fig. 3. Combinatorial TFBS analysis result. The user can see the combinatorial TFBS analysis result. The lists are sorted by p-values.

To this end, FCAnalyzer provides the processes to select significant TFBSs sorted by the p-value of Fisher's exact test, and then to find the significant GO terms in putative target genes.

TFBS-TF matrix constructor

With the FCAnalyzer analysis tool, the user is provided with information about TFBS, a matrix and related TFs including a description and binding information. However, there are numerous TFBSs in the genes of a given cluster. It is possible the binding sub-network of several TFBSs and TFs of interest may not be accessed. Here, we have developed the TFBS-TF matrix constructor in such a way that it facilitates easy and clear analysis of TFs and TFBS information. This module creates a TFBS-TF binding matrix for the analysis in an interactive, web-linked table format (Fig. 4–6). The user can view multiple binding interactions in the context of the TFBS or the TF (Fig. 5). This program also supports interactive TFBS filtering of low occurrences and the selection of TFBSs for each cluster (Fig. 4). This procedure could also help to remove false discoveries.

Functional annotation of gene clusters

In the cluster analysis of gene expression profiles, functional annotation of the cluster is the major method used for identifying functions of co-expressed genes. There are various ways to perform the analysis in relation to the biological pathway, gene annotation and disease information (Dennis *et al*., 2003; Chung *et al*., 2004; Dai *et al*., 2004). A

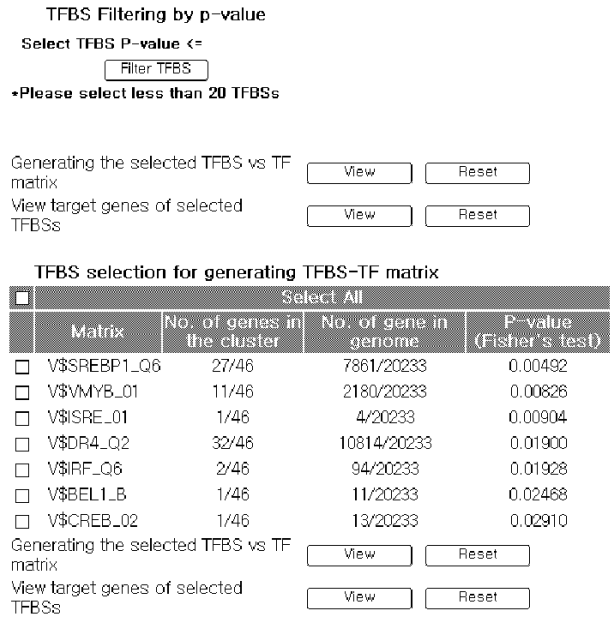


Fig. 4. Web interface of interactive TFBS analysis. The user can filter out non-significant TFBS candidates and view the target genes of TF which binds to selected significant TFBSs.

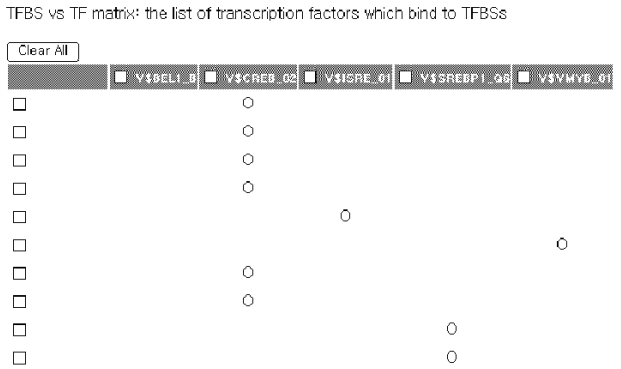


Fig. 5. TFBS-TF Matrix view of selected cluster. TFBS-TF matrix constructor facilitates the analysis of TFBS-TF binding interactions with an easy interface and web-linked TF and TFBS information.

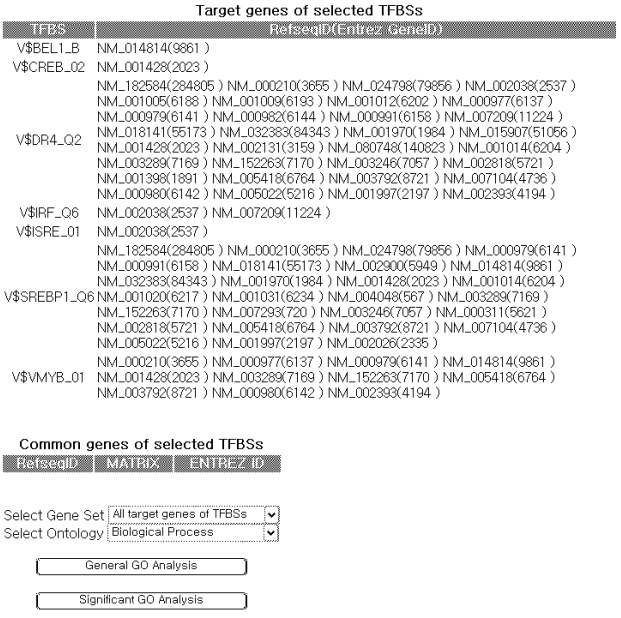


Fig. 6. Target genes of selected cluster. The user can see the target genes of selected TFBSs. General GO analysis and significant GO term analysis are also available for selected target genes.

FCAnalyzer provides the user with the functional analysis of the gene cluster of interest with GO information (Fig. 7 and 8).

FCAnalyzer provides GO term analysis of the co-expressed genes to recognize enrichment GO terms in a cluster. We used the tool GO::TermFinder (Boyle *et al*., 2004) to locate statistically significant over-represented GO terms, as this tool calculates p-values relative to their hypergeometric distribution, and also performs a multiple comparison correction. Therefore, GO term analysis of the clustered genes is informative, and provides a robust

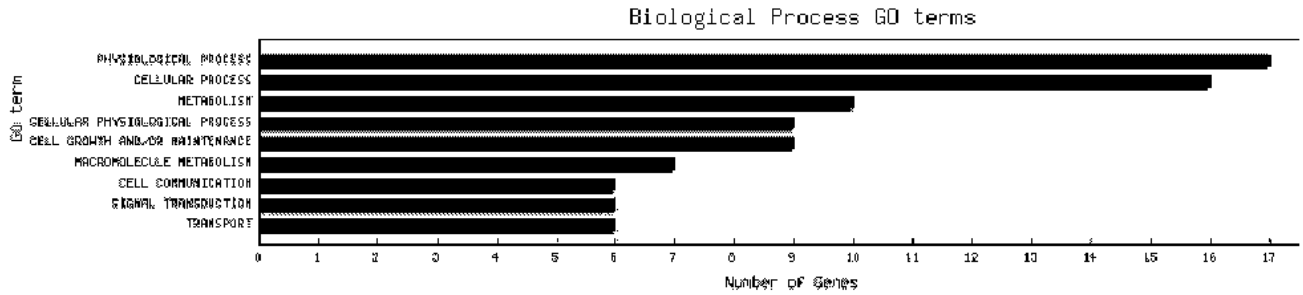
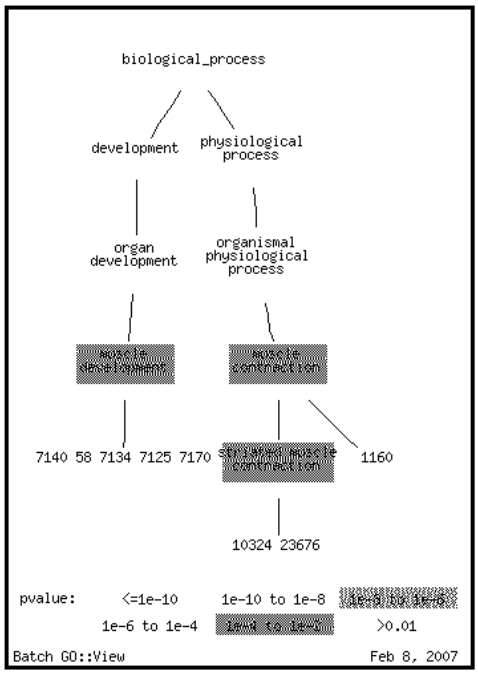


Fig. 7. Gene Ontology view of selected cluster. The user can select one of three categories for general or significant GO term analysis of a selected cluster in terms of biological process, molecular function or cellular component. The result chart for the general GO term analysis is shown.

Biological Process Terms



Result Table

Terms from the Process Ontology with p-value as good or better than 0.05

Gene Ontology term	Cluster frequency	Genome frequency of use	Corrected P-value	FDR	False Positives	Genes annotated to the term
muscle development	5 out of 77 genes, 6.5%	38 out of 11889 genes, 0.3%	0.00031	0.00%	0.00	7140, 58, 7134, 7125, 7170
muscle contraction	3 out of 77 genes, 3.9%	9 out of 11889 genes, 0.1%	0.00160	0.00%	0.00	10324, 1160, 23676
striated muscle contraction	2 out of 77 genes, 2.6%	2 out of 11889 genes, 0.0%	0.00310	0.00%	0.00	10324, 23676

Fig. 8. Enrichment analysis of Gene Ontology terms of selected cluster. The table shows representative TFBS sorted by the p-value. The user can view the tree view of significant GO terms and html output format of the result.

statistical arguement as the output data can be easily and intuitively viewed and explored in a web browser.

The user can select ontologies of interest for viewing functional annotation of any selected cluster in terms of biological process, molecular function or cellular component. For an easier method to examine dominant terms in the cluster, the user can also set the parameters to filter any terms of low occurrence. All figures and result files in a tab-delimited text file format can be downloaded from the download menu of each cluster.

Results

FCAnalyzer supports graphical views and interactive information retrievals to provide a user friendly interface. This web based system provides researchers with the useful features of information databases, sequential clustering analysis of gene expression data and cluster information for functional analysis. To view all of the information on a page, the user can choose to hide or show information and the analysis menu. We anticipate that these user friendly application interfaces will facilitate the analysis of gene expression patterns in the field of transcriptomics.

Databases

FCAnalyzer provides gene expression clustering information and TFBS clustering information of gene expression profiles. The system offers the description and binding information of TF and TFBS for regulatory analysis. All of the information is implemented in a web-linked table format in order to permit compatibility with interconnected information. The annotation information is mainly retrieved from the TRANSFAC database, DAVID, and the Refseq database.

Clustering analysis

FCAnalyzer includes sequential clustering analysis procedures with K-means clustering of gene expression data and subsequent hierarchical clustering of predicted transcription factor binding sites. To attain superior clustering results, users are free to choose other clustering methods (e.g. hierarchical clustering) in the pre-analysis module, and FCAnalyser allows the researcher to determine an initial set of cluster centers. The user can view TFBSs with high frequencies in genes that have a similar expression patterns. The TFBSs with high frequency are presented as the putative regulatory elements playing an important role in gene expression regulation. After the two-step clustering analysis, FCAnalyzer produces TFBS clustering images for genes in each cluster, and images for their respective gene expression profiles. All of the information required for clustering analysis is freely available to download.

Exploring the biological relevance of clusters

In this process, the user can explore the biological relevance of genes identified in the clusters. This step is an important progression which functions to organize genes with similar expression patterns into biologically relevant clusters using GO information (Maurer *et al.*, 2005). The user can select ontologies of interest for viewing the functional annotation of the selected cluster. FCAnalyzer supports a visual graph chart for viewing the GO terms in order of frequency. This graphical interface enables the user to understand the functional specificity of genes in the designated cluster. For TF and TFBS interaction analysis, the TFBS-TF matrix constructor creates TFBS-TF binding matrices from TFBS-TF databases for easy viewing of regulatory interactions.

Searching for GO terms and TFBS analysis to generate a query gene list

FCAnalyzer provides GO terms and TFBS analysis to generate a query gene list without expression values. The user can search for the GO and TFBS information for genes selected after pre-analysis or other processes. They may also perform further analysis such as the analysis of target genes against significant TFBSs.

System performance evaluation

For the evaluation of system performance, we used muscle-specific genes derived from the T-STAG database (Gupta *et al.*, 2005). T-STAG is a database for analyzing tissue- and tumor-specific expression patterns in the human and mouse transcriptomes. We followed the default parameters of T-STAG to obtain genes showing muscle-specific expression. Of these, 91 Refseq mRNA records were submitted to the statistical analysis module of FCAnalyzer. We applied Fisher's exact test to the muscle gene sets. Through this statistical analysis, we then predicted the overrepresented TFBS results in muscle-specific genes. Next, we evaluated the analysis results using literature mining. We found that most of the putative muscle-specific regulators have been reported. For example, the well-known SMAD and MAZ binding sites are ranked as high as second and third in the predicted result. In case of SMAD, dysfunctional SMAD signaling contributes to abnormal smooth muscle cell proliferation (Yang *et al.*, 2005) and MAZ has been implicated in muscle function (Germain-Desprez *et al.*, 2001). Also, mef2 directly regulates target genes at all stages of muscle development (Sandmann *et al.*, 2006).

Therefore, 11 out of 15 putative TFBS identified here are known to be muscle-specific regulators (Table 1).

Discussion

A number of previously developed cluster analysis tools for the study of microarray gene expression data can produce lists of up to hundreds of genes in groups or clusters of putatively related genes. We have developed FCAnalyzer as a tool to help interpret the biological significance of such clusters by using transcription regulatory elements that are based on predicted TFBSs. One limitation of this tool is that the input data needs to be filtered by some criteria for the enhanced TFBS clustering. Nevertheless, we expect that FCAnalyzer will serve as a useful tool for the exploratory analysis of transcription regulatory mechanisms.

This program is designed to be useful for any researcher who is confronted with data that show putative binding site patterns that are shared by coexpressed genes. It is also expected to provide enhanced verification of the results of analyzing gene expression profiles, by offering TFBS information for genes with similar expression patterns.

Availability and requirements

Project name: FCAnalyzer
Project home page: http://www.ngri.go.kr/cgi-bin/cmams/fcanalyzer.cgi
Operating system(s): Linux
Programming language: PERL, HTML, R language
Other requirements: MySQL
Any restrictions to use by non-academics: License required

List of abbreviations

Transcription factor (TF)
Transcription factor binding site (TFBS)
Gene Ontology (GO)
Position weight matrix (PWM)

Acknowledgements

The authors would like to thank Sung-Hoon Lee, Jeong-Ho Cha, Chang-Bum Hong and Dong-Jun Kim for computer system support and helpful comments. This study was supported by an intramural fund from the National Institute of Health, Korea.

References

Ashburner, M., Ball, C. A. *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25-29.

Boyle, E. I., Weng, S. *et al.* (2004). GO:TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* 20, 3710-3715.

Chung, H. J., Kim, M. *et al.* (2004). ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics. *Nucleic Acids. Res.* 32(Web Server issue), W460-W464.

Dai, H., Tian, B. *et al.* (2004). Dynamic integration of gene annotation and its application to microarray analysis. J. Bioinform. *Comput. Biol.* 1, 627-645.

Dennis, G., Jr., Sherman, B. T. *et al.* (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome. Biol.* 4, P3.

Eisen, M. B., Spellman, P. T. *et al.* (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863-14868.

Germain-Desprez, D., Brun, T. *et al.* (2001). The SMN genes are subject to transcriptional regulation during cellular differentiation. *Gene.* 279, 109-117.

Gupta, S., Vingron, M. *et al.* (2005). T-STAG: resource and web-interface for tissue-specific transcripts and genes. *Nucleic Acid. Res.* 33(Web Server issue), W654-W658.

Jenuth, J. P. (2000). The NCBI. Publicly available tools and resources on the Web. *Methods Mol. Biol.* 132, 301-312.

Kasturi, J. and Acharya, R. (2005). Clustering of diverse genomic data using information fusion. *Bioinformatics* 21, 423-429.

Kellis, M., Patterson, N. *et al.* (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241-254.

Kim, J., Seo, J. *et al.* (2005). TFExplorer: integrated analysis database for predicted transcription regulatory elements. *Bioinformatics* 21, 548-550.

Knuppel, R., Dietze, P. *et al.* (1994). TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. *J. Comput. Biol.* 1, 191-198.

Liu, Y., Wei, L. *et al.* (2004). A suite of web-based programs to search for transcriptional regulatory motifs. *Nucleic Acids Res.* 32(Web Server issue), W204-W207.

Lobenhofer, E. K., Bushel, P. R. *et al.* (2001). Progress in the application of DNA microarrays. *Environ Health Perspect* 109, 881-891.

Martin, D., Brun, C. *et al.* (2004). GOToolBox: functional analysis

of gene datasets based on Gene Ontology. *Genome Biol.* 5, R101.

Maurer, M., Molidor, R. *et al.* (2005). MARS: microarray analysis, retrieval, and storage system. *BMC Bioinformatics* 6, 101.

Murakami, K., Kojima, T. *et al.* (2004). Assessment of clusters of transcription factor binding sites in relationship to human promoter, CpG islands and gene expression. *BMC Genomics* 5, 16.

Prestridge, D. S. (1995). Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol.Biol.* 249, 923-932.

Pruitt, K. D., Tatusova, T. *et al.* (2003). NCBI Reference Sequence project: update and current status. *Nucleic Acids Res.* 31, 34-37.

Roth, F. P., Hughes, J. D. *et al.* (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* 16, 939-945.

Sandelin, A., Wasserman, W. W. *et al.* (2004). ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.*32(Web Server issue), W249-W252.

Sandmann, T., Jensen, L. J. *et al.* (2006). A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev. Cell* 10, 797-807.

Sausville, E. A. and Holbeck, S. L. (2004). Transcription profiling of gene expression in drug discovery and development: the NCI experience. *Eur. J. Cancer* 40, 2544-2549.

Shamir, R., Maron-Katz, A. *et al.* (2005). EXPANDER--an integrative program suite for microarray data analysis. *BMC Bioinformatics* 6, 232.

Sinha, S. and Tompa, M. (2003). YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* 31, 3586-3588.

Solovyev, V. and Salamov, A. (1997). The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 5, 294-302.

Sosinsky, A., Bonin, C. P. *et al.* (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.

Tavazoie, S., Hughes, J. D. *et al.* (1999). Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281-285.

Villard, J. (2004). Transcription regulation and human diseases. *Swiss Med. Wkly.* 134, 571-579.

Walsh, B. and Henderson, D. (2004). Microarrays and beyond: what potential do current and future genomics tools have for breeders? *J. Anim. Sci.* 82 E-Suppl, E292-E299.

Yang, X., Long, L. *et al.* (2005). Dysfunctional Smad signaling contributes to abnormal smooth muscle cell proliferation in familial pulmonary arterial hypertension. *Circ. Res.* 96, 1053-1063.