cGRNexp: a Web Platform for Building Combinatorial Gene Regulation Networks based on user-uploaded gene expression datasets

Huayong Xu[1*], Hui Yu [2,3*], Kang Tu[2], Qianqian Shi[3], Chaochun Wei [1,2],Yuanyuan Li[†2], Yixue Li[†1,2,3]
1 School of Life Sciences and Biotechnology, Shanghai Jiao Tong University,100 Dongchuan Road, Shanhgai 200240, P.R.China
2 Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai 200235, P.R.China
3 Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, P.R.China
†corresponding authors E-MAIL: yyli@scbit.org, yxli@scbit.org

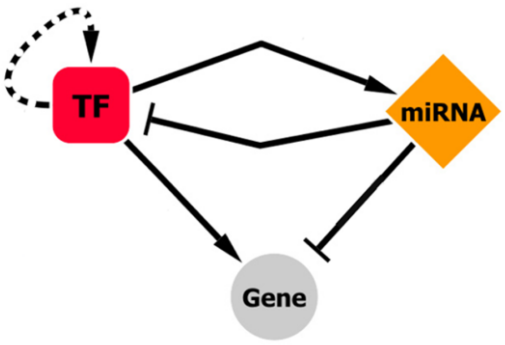
***Abstract*—While we witness rapid progresses in development of methodologies/algorithms for constructing and analyzing the combinatorial regulation network which includes both TF regulators and miRNA regulators, we find a lack of tools or servers available for facilitating related works. A web service is especially needed that allows user to upload their own expression datasets and mine the combinatorial gene reglatory networks regarding the particular experimental context. Herein we report cGRNexp, a web platform for building combinatorial gene regulation networks based on user-uploaded gene expression datasets. In cGRNexp, we deposit three types of sequence-matching-based regulatory relationships and implement two functional modules for processing microRNA-perturbed gene expression datasets and parallel miRNA/mRNA expression datasets. With the microarrays and next-generation sequencing platforms being increasingly accessible, a large amount of miRNA or mRNA expression datasets will be attainable in the near future, and thus, our web platform cGRNexp will be very useful for helping people mine the conditional combinatorial regulatory networks from their own expression datasets. cGRNexp is accessible at http://www.scbit.org/cgrnexp/.**

***Keywords*—combinatorial gene regulatory network; trascription factor; microRNA; webserver**

I. BACKGROUND

Metazoan genomes are characterized with two major types of gene expression regulations: transcription factor (TF) regulation at the transcription level and microRNA regulation at the post-transcription level. These two types of gene regulations interact intimately with each other to govern many important biological processes (Martinez and Walhout 2009; Shalgi, Brosh et al. 2009) and confer robustness against system noise (Herranz and Cohen 2010). In the past few years, great efforts have been devoted to delineate and characterize TF and miRNA-mediated combinatorial gene regulation networks (Fig. 1). In some works (Shalgi, Lieber et al. 2007; Re, Cora et al. 2009; Guo, Sun et al. 2010), complementarity between seed-sequences of regulators and potential targets were exploited to infer putative regulator-to-target relationships and a combinatorial regulation network was constucted through combining heterogeneous regulation relationships. In other works (Essaghir, Toffalini et al. 2010; Meng, Chen et al. 2011; Naeem, Kuffner et al. 2011), gene expression data were integrated with the sequence-matching information and conditional combinatorial regulation networks were mapped .

Despite these methodologies/algorithms for combinatorial regulatory network modeling, we find a lack of tools or servers available for facilitating related works. Till now, only some resources that are non-relevant to expression data are open to public, of which MIR@NT@N (Le Bechec, Portales-Casamar et al. 2011) (http://maia.uni.lu/mironton.php) is a recent example. As a matter of fact, biomedical researchers are more interested in constructing conditional combinatorial regulation networks in which regulations specific to a particular biological condition are extracted through making use of conditional gene expression data. A web service allowing users to upload their own expression datasets and extract a conditional combinatorial regulation network is apparently more useful. TFactS (Essaghir, Toffalini et al. 2010) (www.tfacts.org) is a web-tool of this kind, but it addresses the problem in a highly simplified fashion – differentially expressed genes, instead of the full expression dataset, are read in, and only the active transcription factors, rather than a comprehensive combinatorial regulation network, are returned.



978-1-4673-4398-5/12/$31.00 ©2012 IEEE

Figure 1. Basic regulation types in TF and miRNA-mediated combinatorial gene regulation networks. TF and miRNA are two different types of gene expression regulators, both being able to regulate protein-coding genes (termed "Gene" in figure). TF regulates protein-coding genes or miRNA genes at the DNA-sequence level (TF→gene and TF→miRNA, common arrow in figure), while miRNA regulates protein-coding genes at the mRNA level（miRNA→gene，blunt arrow in figure）。In the figure, the same node "Gene" is used to signify both the DNA level and the mRNA level. The blunt arrows diriving from miRNA to "TF" or "Gene" are in nature of the same type, together termed the miRNA→gene regulation。Despite the existence of self-regulation of TF (dashed arrow), for simplicity it is often ignored in mathmatical modelling. In all, the basic regulation types in TF and miRNA-mediated combinatorial gene regulation networks are classified to three types: TF→miRNA, miRNA→gene, and TF→gene.

Our group has developed two algorithms for mapping conditional combinatorial gene regulation networks based on sequence-complentarity information and expression data. In the first algorithm (Tu, Yu et al. 2009), we utilize the MPGE (MicroRNA-Perturbed Gene Expression) datasets, in which a particular miRNA of interest is over-expressed or knocked-down, and large-scale mRNA expression levels are measured before and after the miRNA perturbation. With a MPGE dataset as well as sequence-based TF-gene and miRNA-gene putative regulatory relationships, we are able to build a two-layer regulatory network in which transcription factors (TFs) function as important mediators of miRNA-initiated regulatory effects. In the second algorithm (Yu, Tu et al. 2012), we utilize parallel miRNA and mRNA expression datasets in order to mine a comprehensive combinatorial gene regulatory network in a specific biological context. With the parallel miRNA and mRNA expression datasets as well as sequence-based TF-miRNA, TF-gene, and miRNA-gene putative regulatory relationships, we are able to construct a comprehensive combinatorial gene regulatory network which cover three types of regulations.

Having been proved efficient in the previous studies, our combinatorial network construction algorithms are expected to benefit peer researchers in building their own conditional combinatorial gene regulatory networks. With the microarrays and next-generation sequencing platforms being increasingly accessible, many more MPGE datasets and parallel expression datasets will be attainable in the near future, and thus, the field is in dire need of a web tool providing the services of building conditional combinatorial regulation networks based on user-uploaded expression datasets. This is why we developed cGRNexp (http://www.scbit.org/cgrnexp), a web platform for building combinatorial gene regulation networks based on user-uploaded gene expression datasets. At cGRNexp, users can employ our published algorithms to construct conditional combinatorial regulation networks based on their own gene expression datasets and make further analyses. We believe cGRNexp will benefit peer researchers and help advance studies of conditional combinatorial gene regulatory networks.

II. WEB SERVER IMPLEMENTATION

The cGRNexp is a freely accessible and attainable tool through the World Wide Web offering the service for modeling combinatorial gene regulatory networks based on built-in regulation libraries and user-uploaded expression datasets. The web server is designed as a PHP+R framework (Fig. 2), with the PHP modules in charge of communicating with web users and internal R modules performing the major calculation. Two functional main modules are included in cGRNexp, namely Mod_MPGE and Mod_Parallel. Mod_MPGE works on MPGE datasets and returns a combinatorial gene regulation network covering two types of regulations (miRNA-gene and TF-gene) (Tu, Yu et al. 2009); Mod_Parallel works on parallel miRNA/mRNA expression datasets and returns a combinatorial gene regulation network covering three types of regulations (miRNA-gene, TF-gene, and TF-miRNA) (Yu, Tu et al. 2012).

The user-friendly web interface allows users to use R modules without having any dealings with the R environment. The only requirement of utilizing cGRNexp is a web browser that can access the internet. Users submit their own expression datasets and set the required parameters at the web interface before submitting their jobs. When the job is finished, the user can view and download the analysis results with a HTML formatted web page at a URL sent to the user's email.

On each page of cGRNexp, a navigation bar is placed on the top where users can click to jump among five tabs. The 'Home' page gives general information on cGRNexp; the 'Mod_MPGE' and the 'Mod_Parallel' pages are where users can submit their calculation jobs; the 'Download' page offers some resources to be downloaded; finally, the 'Help' page includes comprehensive helping information to ease using cGRNexp (Fig. 3).

III. DATA LIBRARIES

Three data libraries, TF2gene, TF2miR, and miR2gene, are deposited in cGRNexp as built-in data components required for calculation. The TF2gene and TF2miR libaries, including putative TF→gene and TF→miRNA regulation relationships based on matching of TF binding sites, were developed mainly from the 'tfbsConsSites.txt' file and the 'tfbsConsFactors.txt' file from UCSC hg19, the results of scanning human genome for human/mouse/rat conserved TF binding sites (http://genome.ucsc.edu/cgi-bin/hgTables). The miR2gene library, developed from orginal dataset from starBase (Yang, Li et al. 2011) (http://starbase.sysu.edu.cn/), includes putative miRNA→gene regulation relationships mapped from CLIP-Seq and Degradome-Seq data. These data libraries will be regularly updated to keep up with the respective source files, and are open to academic users for free download. The version of the data libraries could be found on the main module pages and also on the download pages. For further information on processing and accessing these data libraries, please refer to help pages of cGRNexp.

Wherever microRNA names occurs, we processed the original data files in order to uniform different miRNA names representing identical miRNA transcripts that were transcribed from different genome coordinates (for example, 'hsa-let-7a-1', 'hsa-let-7a-2', and 'hsa-let-7a-3' are uniformed to 'hsa-let-7a'). This rule applies to tf2miR and mir2gene libraries. We require users to uniform miRNA names in their expression dataset in a similar fashion.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

IV. Functional Modules

In Mod_MPGE, we work on the MPGE (MicroRNA-Perturbed Gene Expression) dataset to build a two-layer miRNA-driven combinatorial regulatory network. In a MPGE experiment, miRNA is transfected into a particular cell line, and a certain time period (usually 12h or 24h) later, the mRNA levels in the miRNA-transfected cells are measured and compared with pre-transfection mRNA levels. The server first determine the miRNA's overall degradation-inducing ability using one-sided Kolmogorov–Smirnov (K–S) test and identify its putative degraded targets using non-parametric tests. Then it uses the stepwise linear regression model to integrate the MPGE dataset with the seed-sequence-matching-based TF2gene and miR2gene putative regulatory relationships, and as a result identify the active TF mediators of the regulation cascades initiated by the perturbed miRNA. At last, the server determines a subset of regulatory relationships from the TF2gene and miR2gene data libraries which correspond to a two-layer network centered around the perturbed miRNA and the few TF mediators.

In Mod_Parallel, we utilize parallel miRNA and mRNA expression datasets in order to map a combinatorial gene regulatory network in a specific biological context. The inputs to Mod_Parallel include the three types of putative regulatory relationships (TF2gene, miR2gene, and TF2miR), and the parallel microRNA (miRNA) and mRNA expression datasets measured in a same series of experimental conditions. In addition to mapping a comprehensive TF-and-miRNA involved combinatorial regulatory network, we also pinpoint the important vertices and edges in the resultant combinatorial network in terms of degree ranking or betweenness ranking. Besides, we also seek 'co-regulating regulator pairs' in the combinatorial regulatory network. For this task, we scrutinize all regulator pairs by testing the significance of their co-regulating common targets. We return the co-regulating regulator pairs that pass the 'fisher exact test p-value cutoff'. Finally we perform the 'triple-vertex motif analysis'. Theoretically there are a total of eighteen triple-vertex regulatory motifs, defined as closed triple-vertex regulatory circuits that involve at least a miRNA and a TF, and they can be classified into 'feed-forward loops' (FFLs) and 'feed-backward-loops' (FBLs) by considering the ways of directed regulations being connected. We count in the resultant combinatorial network the occurrences of all possible triple-vertex motifs, and estimate the corresponding p-values through counting the counterpart occurrences N times in randomly shuffled networks.

V. Examples/Case Studies

A. A case study of Mod_MPGE

We tested MPGE with the 24-hour MPGE dataset of hsa-miR-1 (sample dataset at Mod_MPGE; obtained from http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS1858). This dataset records the expression log ratios of 20,127 protein-coding genes obtained by comparing the expression profiles of HeLa cells after and before transfection of hsa-miR-1. It took our web server around 5 minutes to finish the whole calculation.

The major results include three parts: 1) the degraded targets of the core miRNA (hsa-miR-1 in this case) and a PP-plot for giving further information on the miRNA's degraded target analysis; 2) the active TF mediators; 3) the two-layer regulatory network with the active TFs mediating the miRNA-initiated regulatory effects. Most results are displayed as tab-delimited text tables with the gene identities and relevant statistics shown explicitly (http://www.scbit.org/cgrnexp/doR_target_result.php?jobID=821338391035). As R is not an excellent tool for displaying large-scale networks, we suggest the user to view the resultant network by loading the original text file 'network.edge.txt' from the report page to an external graphical tool, such as CytoScape (Smoot, Ono et al. 2011).

B. A case study of Mod_Parallel

Two parallel cancer gene expression datasets, one for miRNA and another for mRNA, were downloaded from CellMiner (http://discover.nci.nih.gov/cellminer/loadDownload.do). These two datasets were generated to study the 60 human cancer cell lines of the NCI-60 program using the 41,000-probe Agilent Whole Human Genome Oligo Microarray and the 15,000-feature Agilent Human microRNA Microarray V2.

The original miRNA expression dataset included only 365 microRNAs with detectable expression (Liu, D'Andrade et al. 2010). After miRNA names were uniformed and corresponding data lines averaged, this dataset covered 266 human miRNAs. The mRNA expression dataset downloaded immediately had more than 41K data rows, with 40,155 rows having entrez GeneID identification. After removing non-identified data rows and combining different rows for identical genes, we obtained a mature mRNA expression dataset for 21,319 human protein-coding genes. These two parallel expression datasets, one for miRNAs and another for mRNAs, are available at Mod_Parallel as sample datasets. It took the our web server around 60 minutes to finish the whole calculation of these two datasets.

The major results include four parts: 1) the resultant combinatorial gene regulatory network, shown as the text edge file; 2) the vertices/edges of the network ranked by topological features; 3) significantly co-regulating regulator pairs; 4) significance of recurrence of 18 triple-vertex regulatory motifs and all instances of the appeared motifs. For the full results, see http://www.scbit.org/cgrnexp/doR_network_result.php?jobID=931338180093.

VI. Conclusion

It is of fundamental importance to delineate and characterize combinatorial gene regulatory networks that correspond to particular cellular contexts. A successful strategy addressing this problem is to integrate the sequence-matching-based regulatory relationships with conditional gene expression datasets through regression models. Herein we report cGRNexp, a web platform for modeling combinatorial

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

gene regulation networks based on user-uploaded gene expression datasets. Currently cGRNexp implements two functional modules that work on MPGE dataset and parallel expression datasets respectively. In the future, we will update the sequence-matching-based regulatory relationships regularly and incorporate more functional modules targeted at mapping combinatorial regulatory networks. With the microarrays and next-generation sequencing platforms being increasingly accessible, a large amount of miRNA or mRNA expression datasets will be attainable, and thus, our web platform cGRNexp will be very useful for helping people mine the conditional combinatorial regulatory networks from their own expression datasets.

ACKNOWLEDGEMENT

We would like to thank Yijie Wang and Junzhe Mao from Shanghai High School, Shanghai, China for their help on updating data libraries.

This work was supported by grants from National key basic research program (funding numbers: 2011CB910204, 2012CB316501, 2011CB910200, 2010CB912702); Chinese Academy of Sciences (Innovation funding KSCX2-YW-R-112, KSCX2-EW-R-04); National Natural Science Foundation of China (31000380 to HY, 31171268 to LYY, 90913009 to YXL).

REFERENCES

[1] Essaghir, A., F. Toffalini, et al. (2010). "Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data." Nucleic Acids Res 38(11): e120.

[2] Guo, A. Y., J. Sun, et al. (2010). "A novel microRNA and transcription factor mediated regulatory network in schizophrenia." BMC Syst Biol 4: 10.

[3] Herranz, H. and S. M. Cohen (2010). "MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems." Genes Dev 24(13): 1339-44.

[4] Le Bechec, A., E. Portales-Casamar, et al. (2011). "MIR@NT@N: a framework integrating transcription factors, microRNAs and their targets to identify sub-network motifs in a meta-regulation network model." BMC Bioinformatics 12: 67.

[5] Liu, H., P. D'Andrade, et al. (2010). "mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities." Mol Cancer Ther 9(5): 1080-91.

[6] Martinez, N. J. and A. J. Walhout (2009). "The interplay between transcription factors and microRNAs in genome-scale regulatory networks." Bioessays 31(4): 435-45.

[7] Meng, J., H.-I. Chen, et al. (2011). Uncover cooperative gene regulations by microRNAs and transcription factors in glioblastoma using a nonnegative hybrid factor model Proc. Int. Conf. Acoustics, Speech and Signal Processing 2011, Prague, Czech Republic.

[8] Naeem, H., R. Kuffner, et al. (2011). "MIRTFnet: analysis of miRNA regulated transcription factors." PLoS One 6(8): e22519.

[9] Re, A., D. Cora, et al. (2009). "Genome-wide survey of microRNA-transcription factor feed-forward regulatory circuits in human." Mol Biosyst 5(8): 854-67.

[10] Shalgi, R., R. Brosh, et al. (2009). "Coupling transcriptional and post-transcriptional miRNA regulation in the control of cell fate." Aging (Albany NY) 1(9): 762-70.

[11] Shalgi, R., D. Lieber, et al. (2007). "Global and local architecture of the mammalian microRNA-transcription factor regulatory network." PLoS Comput Biol 3(7): e131.

[12] Smoot, M. E., K. Ono, et al. (2011). "Cytoscape 2.8: new features for data integration and network visualization." Bioinformatics 27(3): 431-2.

[13] Tu, K., H. Yu, et al. (2009). "Combinatorial network of primary and secondary microRNA-driven regulatory mechanisms." Nucleic Acids Res 37(18): 5969-80.

[14] Yang, J. H., J. H. Li, et al. (2011). "starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data." Nucleic Acids Res 39(Database issue): D202-9.

[15] Yu, H., K. Tu, et al. (2012). "Combinatorial Network of Transcriptional Regulation and microRNA Regulation in Human Cancer." BMC Systems Biology: accepted.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

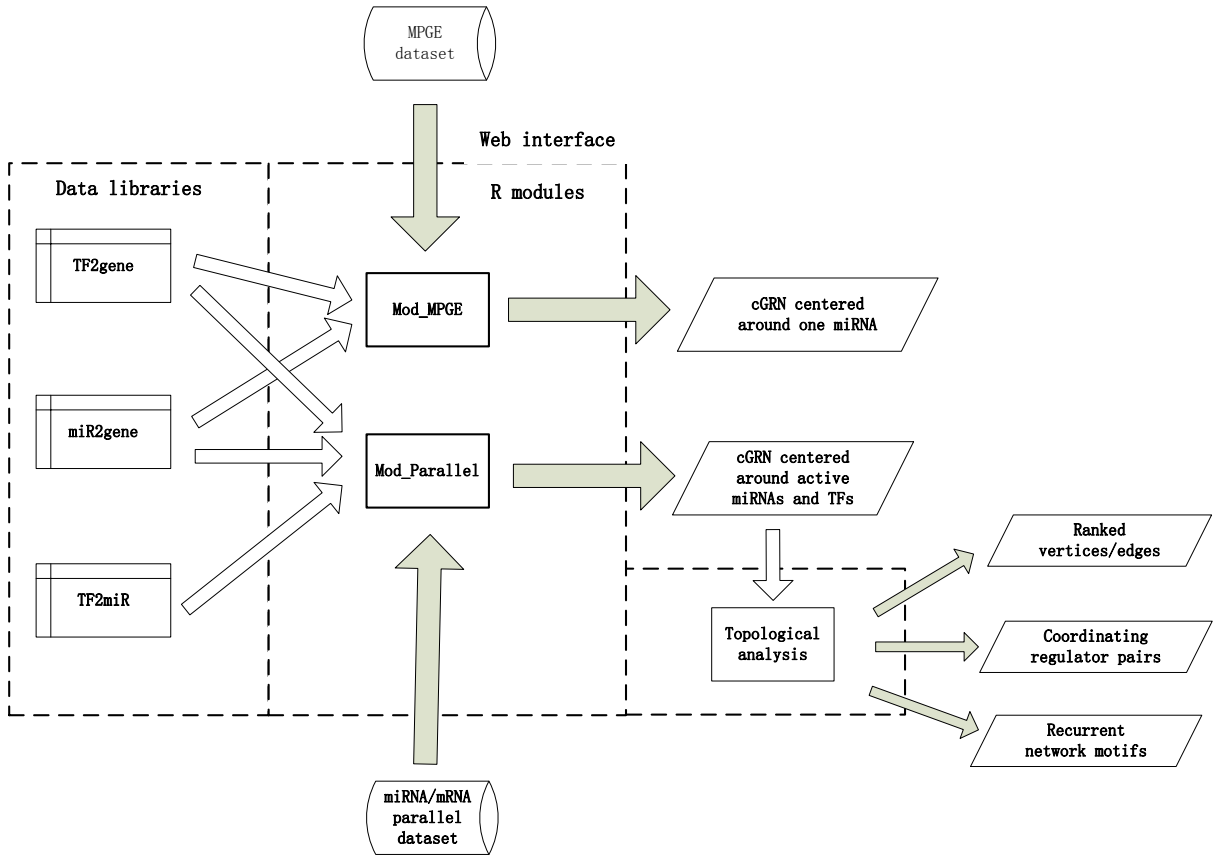


Figure 2. Schematic framework of cGRNexp. The R statistical environment is employed to carry out the two functional modules Mod_MPGE and Mod_Parallel based on built-in regulator-to-target libraries (TF2gene, miR2gene, and TF2miR) and user-uploaded expression datasets; the PHP technology is utilized to deal with the data input and output.

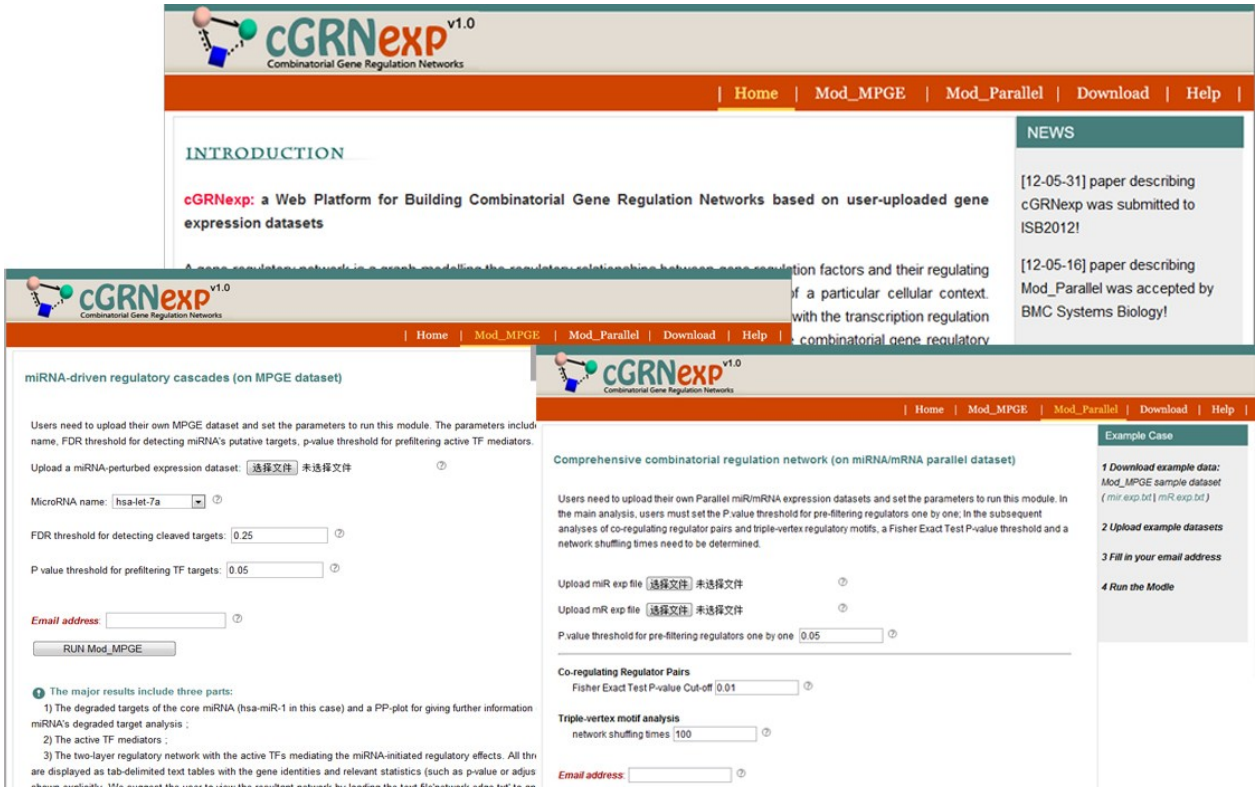


Figure 3. Some screenshots of cGRNexp web pages

978-1-4673-4398-5/12/$31.00 ©2012 IEEE