Published in final edited form as:
Int J Comput Biol Drug Des. 2009 ; 2(1): 1–20.

Identification of condition-specific regulatory modules through multi-level motif and mRNA expression analysis

Li Chen, **Jianhua Xuan**, and **Yue Wang**
Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA

Li Chen: lchen06@vt.edu; Jianhua Xuan: xuan@vt.edu; Yue Wang: yuewang@vt.edu

Eric P. Hoffman
Research Center for Genetic Medicine, Children's National Medical Center, Washington, DC 20010, USA

Eric P. Hoffman: ehoffman@cnmcresearch.org

Rebecca B. Riggins and **Robert Clarke**
Department of Oncology and Physiology & Biophysics, Georgetown University, School of Medicine, Washington, DC 20057, USA

Rebecca B. Riggins: rbr7@georgetown.edu; Robert Clarke: clarker@georgetown.edu

Abstract

Many computational methods for identification of transcription regulatory modules often result in many false positives in practice due to noise sources of binding information and gene expression profiling data. In this paper, we propose a multi-level strategy for condition-specific gene regulatory module identification by integrating motif binding information and gene expression data through support vector regression and significant analysis. We have demonstrated the feasibility of the proposed method on a yeast cell cycle data set. The study on a breast cancer microarray data set shows that it can successfully identify the significant and reliable regulatory modules associated with breast cancer.

Keywords

transcription regulatory module; motif enrichment analysis; SVR; support vector regression; statistical significance analysis; multi-level regulator identification

1 Introduction

The identification of regulatory modules is one of the important yet challenging problems towards understanding the underlying mechanisms of biological processes, especially for pinning down the pathways causing cancers. In the transcriptional level, a regulatory module is defined as a set of genes controlled by one or several Transcription Factors (TFs) in a condition-specific manner (Segal et al., 2003). TFs can either activate or inhibit gene expression through a short highly conserved DNA sequence in the gene promoter (or

Correspondence to: Jianhua Xuan, xuan@vt.edu.

Reference to this paper should be made as follows: Chen, L., Xuan, J., Wang, Y., Hoffman, E.P., Riggins, R.B. and Clarke, R. (2009) 'Identification of condition-specific regulatory modules through multi-level motif and mRNA expression analysis', *Int. J. Computational Biology and Drug Design*, Vol.

upstream) region, i.e., a Transcription Factor Binding Site (TFBS) or binding motif. In higher eukaryotes, TFBSs are often organised in clusters called *cis*-Regulatory Modules (CRMs). Many computational methods have been developed to facilitate the identification of CRMs from either DNA sequence data or gene expression data. Expression-based methods (Segal et al., 2003; Ihmels et al., 2004; Wang et al., 2005) take advantage of gene expression data but lacking of sequence binding constraints. Sequence-based module discovery algorithms, such as CisModule (Zhou and Wong, 2004), CREME (Sharan et al., 2003) and ModuleSearch (Aerts et al., 2003), analyse the promoter regions of a set of co-regulated genes to identify overrepresented motif combinations. A major limitation of the sequence-based methods lies in that the methods do not consider the condition-specific nature of regulatory modules, i.e., ignoring the relationship between binding strengths and gene expression levels as described next.

As we know, a living cell is a dynamic system in which gene activities and interactions exhibit temporal patterns and spatial compartmentalisation (Qi and Ge, 2006). Recently, several studies have shown that binding of TFs depends not only on their affinity to the binding sites, but also on their expression levels (Lee et al., 2002; Segal et al., 2008). This means that a transcription factor may play different regulation roles to its downstream target genes or even has different downstream targets under different conditions (Lee et al., 2002). Motivated by this understanding, many computational algorithms were proposed to discover condition-specific regulatory modules by integrating gene expression profiles and motif information. For example, Ruan and Zhang (2006) proposed a bi-dimensional regression tree approach to model gene expression regulation. Das et al. (2006) utilised linear splines to correlate the binding strengths of motifs with the expression levels. Segal et al. (2008) proposed a thermodynamic model to predict expression patterns from regulatory sequence in *Drosophila* segmentation.

Although these methods have achieved some degree of success, high false-positive prediction rate is still a major problem mainly due to the noises in motif information and gene expression data. To reduce the false-positive rate, in this paper we propose a novel method, namely multi-level regulatory module identification, to help find significant and stable regulatory modules. The method is strengthened through several ways:

- Support Vector Regression (SVR) is utilised to formulate the relationship between motif binding strengths and gene expression levels, aiming to improve the noise-tolerance capability
- a significance analysis procedure is designed to help identify statistically significant regulatory modules
- a multi-level analysis strategy is developed to further reduce the false-positive rate for reliable regulatory module identification.

We have applied our proposed method on a yeast cell cycle microarray data set and a breast cancer microarray data set to identify condition-specific regulatory modules. The experimental results on the yeast cell cycle data set demonstrate the effectiveness of the proposed approach in identifying cell cycle-related cooperative regulators and their target genes. The experimental results on the breast cancer data set further show that the proposed method can be used to identify condition-specific regulatory modules in breast cancer development, which may have important implications to understanding the pathways associated with breast cancer.

2 Methodology

2.1 Sequence analysis for motif binding strength

ChIP-on-chip, also known as genome-wide location analysis, is a technique for isolation and identification of the DNA sequences occupied by specific DNA binding proteins in cells. However, it is not a trivial task to measure the binding strengths for all TFs from ChIP-on-chip experiments due to the limited antibodies available, especially for human studies. An alternative and practical way is to extract motif binding information from the promoter regions of focused genes. Motif is usually represented by a Position Weight Matrix (PWM) that contains log-odds weights for computing a match score between a binding site and an input DNA sequence. Many algorithms have been developed to either *de novo* discover motifs given multiple input sequences (Zhou and Wong, 2004; Bailey et al., 2006) or search the known motifs in a given sequence based on their PWMs (Kel et al., 2003; Chekmenev et al., 2005). Among them, Match^{TM} (Kel et al., 2003) takes DNA sequences as input, searches for potential TF binding sites using a library of PWMs and outputs a list of found potential sites in the sequence. The search algorithm uses two score values: the matrix similarity score (mss) and the core similarity score (css). These two scores measure the quality of a match between the sequence and the matrix, ranging from 0 to 1.0, where 0 denotes no match and 1.0 an exact match. The core of each matrix is defined as the first five most conserved consecutive positions of a matrix.

We assume that the binding strength for a specific transcription factor to its target gene is proportional to the similarity score of its binding site and the number of occurrences of the binding site in the gene promoter region. All human promoter DNA sequences were obtained from the UCSC Genome database (Karolchik et al., 2003) (upstream 5000 bp from the Transcription Start Site (TSS)). With all vertebrate PWMs provided by the TRANSFAC 11.1 Professional Database (Matys et al., 2006), Match^{TM} algorithm is used to generate a gene-motif binding strength matrix $\mathbf{X} = [x_{gm}]$ with the cut offs that minimising the false-positive rate. The rows in the matrix correspond to different genes and columns correspond to different binding sites (or motifs). Each element x_{gm} represents the binding strength of motif m in the promoter region of a gene g, which is mathematically calculated as follows:

$$x_{\mathrm{gm}}=\sum_{i=1}^{N}\frac{1}{2}(\mathrm{mss}_{\mathrm{gmi}}+\mathrm{css}_{\mathrm{gmi}}) \quad (1)$$

where N is the number of occurrences of motif m in the promoter region of gene g; mss$_{gmi}$ and css$_{gmi}$ are the MSS and CSS for motif m and gene g in the ith hit, respectively.

2.2 Support Vector Regression to integrate motif binding strengths and gene expression data

Given a gene set G, its mRNA expression data is represented by a matrix $\mathbf{Y} = [y_{gt}]$, $g \in G$, where each element y_{gt} is the log-ratio of the expression level of gene g in sample t to that of the control sample. Assume M is the active motif set on the gene set G, the corresponding gene-motif matrix is $\mathbf{X} = [x_{gm}]$, $g \in G$, $m \in M$, where x_{gm} is the binding strength of motif m in the promoter region of gene g. The relationship between gene expression levels and motif binding strengths can then be formulated as follows:

$$y_{\mathrm{gt}}=f(x_g)=\sum_{m}a_{\mathrm{mt}}x_{\mathrm{gm}}+b_t \quad (2)$$

where a_{mt} and b_t are the coefficients of the linear regression model. Biologically, the model can be viewed or interpreted as that the log-ratio of gene expression level is the linear combination of log-ratios of Transcription Factor Activities (TFAs) (denoted as a_{mt} in equation (2)) weighted by their binding strengths (i.e., x_{gm}), plus a baseline expression ratio b_t in sample t.

Provided that **X** and **Y** are known, the model is then reduced to a regression problem. Since mRNA expression data and motif binding strength data are noisy, we choose SVR (Smola and Scholkopf, 1998) to solve the coefficients (i.e., a_{mt} and b_t) by using ε-insensitive loss function, with which to ensure the existence of the global minimum and the optimisation of reliable generalisation bound. The ε-insensitive loss function is defined by

$$L_\varepsilon(y_{\mathrm{gt}})=\begin{cases}0, & \text{if } \left|y_{\mathrm{gt}}-\widehat{y}_{\mathrm{gt}}\right|<\varepsilon \\ \left|y_{\mathrm{gt}}-\widehat{y}_{\mathrm{gt}}\right|-\varepsilon, & \text{otherwise}\end{cases}, \qquad (3)$$

where $\widehat{y}_{\mathrm{gt}}$ is the estimated value of expression log-ratio y_{gt}.

The goal of SVR is to find a function f that minimises the loss function while keeping as flat as possible. By introducing slack variables ξ_g and ζ_g^* for soft margin, we can formulate the optimisation problem as follows (Drucker et al., 1997):

$$\begin{aligned}&\text{Minimise}\tfrac{1}{2}\|a\|^2+C\sum_{g\in G}(\xi_g+\xi_g^*),\\ &subjec\ to\begin{cases}y_g-<a,x_g>-b\le\varepsilon+\xi_g\\ <a,x_g>+b-y_g\le\varepsilon+\xi_g^*\\ \xi_g,\xi_g^*\ge 0\end{cases}.\end{aligned} \qquad (4)$$

The constant $C > 0$ determines the trade off between the flatness of f and the amount up to which deviations larger than ε are tolerated.

By further introducing non-negative Lagrangian multipliers a_g and a_g^*, we can formulate the above optimisation problem to be the following equivalent one of maximising the dual Lagrangian function with respect to a_g and a_g^* (Drucker et al., 1997):

$$\begin{aligned}&\text{Maximise}-\tfrac{1}{2}\sum_{g_i,g_j\in G}(a_{g_i}-a_{g_i}^*)(a_{g_j}-a_{g_j}^*)<x_{g_i},x_{g_j}>-\varepsilon\sum_{g\in G}(a_g+a_g^*)+\sum_{g\in G}y_g(a_g-a_g^*)\\ &subject\ to\sum_{g\in G}(a_g-a_g^*)=0,\ a_g,a_g^*\in[0,C].\end{aligned} \qquad (5)$$

By solving the above optimisation problem, we can finally obtain the solution to the regression problem as follows (Drucker et al., 1997):

$$\begin{aligned}&a=\sum_{g\in G}(a_g-a_g^*)x_g,\\ &b=\begin{cases}y_g-\left\langle a,x_g\right\rangle-\varepsilon, & if\ a_g\in(0,C)\\ y_g-\left\langle a,x_g\right\rangle+\varepsilon, & if\ a_g^*\in(0,C)\end{cases},\ and\\ &\widehat{y}_g=f(x_g)=\sum_{g\in G}(a_g-a_g^*)\left\langle x_g,x_g\right\rangle+b.\end{aligned}$$

Finally, there is one important issue that needs to be addressed, that is, how to determine an appropriate motif set for SVR fitting. Due to the large number of motifs under study (typically in a range of 50–500), it is not feasible to consider all possible motif combinations when the order of the motif set increases. In order to reduce the computational complexity, we use a stepwise forward greedy search strategy to find the cooperative motifs in a given gene set. Specifically, starting from each individual motif, we incorporate one more motif into current motif set from the remaining motifs at each step. The selected motif should be the one that can best reduce the fitting error compared to the other motifs. Notice that this procedure is not immune to the overfitting problem, i.e., adding in more motifs will reduce the fitting error mathematically – a result of fitting to the noise rather than the true signal. In order to avoid this problem, we use a modified ε-intensive loss function to determine if a new motif should be added into the current motif set. The modified ε-intensive loss function is defined as follows, by adding a penalty term to equation (3):

$$L(Y,M)=\sum_{g,t}(L_{\varepsilon}(y_{gt}))^2/(1-|M|/|G|)^2. \quad (6)$$

The motif set searching procedure will stop when the following condition is satisfied:

$$L(Y,M_0)<L(Y,(M_0\cup\{m\}),\forall:m\in M_{A}-M_0. \quad (7)$$

where M_A is the whole motif set and M_0 is the current motif set. In our implementation, we allow up to triplet motifs to be searched for, assuming that a regulatory module is only regulated by a small number of TFs.

2.3 Significance analysis of regulatory modules

A significance analysis procedure is designed to test if a selected motif set is statistically significant in regulating a given gene set, aiming to identify active cooperative regulators for a given gene set. The null and alternative hypotheses (H_0 and H_1, respectively) are given as follows:

H_0: *The motif set is not actively regulating a given gene set.*

H_1 : *The motif set is actively regulating a given gene set.*

We design a summary statistic to represent the fitting results shown below:

$$\begin{aligned}&F=\frac{\mathrm{RSS}_0-\mathrm{RSS}_1}{\mathrm{RSS}_1}\\&\mathrm{RSS}_0=\sum_{g,t}(y_{\mathrm{gt}}-\bar{y}_t)^2,\quad \bar{y}_t=\frac{1}{|G|}\sum_{g\in G}y_{\mathrm{gt}},\\&\mathrm{RSS}_1=\sum_{g,t}(y_{\mathrm{gt}}-\widehat{y}_{\mathrm{gt}})^2,\quad \widehat{y}_{\mathrm{gt}}=\sum_{m}a_{\mathrm{mt}}x_{\mathrm{gm}}+b_t\end{aligned} \quad (8)$$

where RSS_0 is the residual sum of squares without motif participation, and RSS_1 is the residual sum of squares with motif participation. The above equation is proportional to the typical F-statistic used in statistics for comparing two models (Lomax, 2007). In order to calculate p -value, we use the permutation method described below to form null distribution. For a given motif set, we randomly select a gene set G_0 with the same size of G from the whole gene population, and then repeat B times to generate the corresponding null statistic score F^{0b}, for $b = 1, 2, \ldots, B$. The p-value can be obtained for each gene set by calculating the probability that a null gene set has a statistic more extreme than the observed statistic. Mathematically, the p-value can be calculated by the following equation:

$$p=\mathrm{Pr}_{H_0}(F^{0b}>F)=\frac{\#\left\{b:F^{0b}>F,b=1,\cdots,B\right\}}{B}. \quad (9)$$

2.4 Multi-level regulatory module identification

Assuming that most genes involved in a regulatory module are co-expressed under a given condition, we can use a clustering method to form the gene set for regression analysis. However, simple gene clustering, on one hand, often results in many false-positives for gene module identification; motif information, on the other hand, is quite noisy and also incomplete due to the current status of limited biological knowledge. Hence, some real cooperative regulators would not be easily revealed, or instead, false-positives would be included based on a fixed gene set. In order to reduce the false-positives, we further postulate that a condition-specific regulatory module and its enriched motifs will appear more and more significantly and stably in different levels of clustering, when the irrelevant genes are gradually eliminated in a coarse-to-fine fashion. Based on this assumption, a multi-level analysis strategy is further developed to enhance the SVR approach described above for reliable regulatory module identification. In particular, a multi-level gene clustering procedure, such as Self-Organising Map (SOM) clustering (Kohonen, 1997), is used to form the gene clusters to gradually reduce the noises in gene expression data and motif information. The flowchart of multi-level analysis procedure is shown in Figure 1 and also can be summarised as follows:

1. Set cluster number $c = 1$ and cluster level $l = 1$. Identify all possible enriched motif sets and calculate their *p*-values on current gene set G through SVR analysis and significance analysis.
2. Increment c by 1 and l by 1. Cluster the gene population into c clusters, denoted as $\left\{G_1^l, G_2^1, \ldots, G_c^l\right\}$.
3. For each gene cluster, discover all possible enriched motif sets and calculate their *p*-values.
4. Repeat Steps 2 and 3 until the following stop criterion is met, that is, the number of genes is less than a threshold t_0 for all gene clusters.
5. Let us use p_M^{lc} to denote the *p*-value of a candidate motif set M at different levels and clusters. Output the significantly and stably enriched motif sets if they satisfy $\min\,(p_M^{\mathrm{lc}}) < p_0^l$, $\forall l$, where p_0^l is the threshold of *p*-value at each level l.
6. Use a voting scheme to determine the gene members of a regulatory module with the enriched motif set M; the voting scheme is described as follows:
   a. Initialise a gene weight vector **w** as 0
   b. Update **w** by the following equation:

$$\forall l, c, \quad w_{G_c^l} = w_{G_c^l} + \sum_{m \in M} X_{G_c^l m}, \quad if\ p_m^{\mathrm{lc}} < p_0^l \ .$$

Finally, the genes whose weights are greater than a threshold w_0 are chosen as the members of a corresponding regulatory module. In our implementation, we set w_0 as the mean of **w** plus one standard deviation.

3 Results

3.1 Data description

We applied the proposed method to two gene expression profiling studies:

- a yeast cell cycle microarray data set (Spellman et al., 1998)

- a breast cancer cell line microarray data set (Creighton et al., 2006).

The yeast cell cycle data set consists of the expression of 6178 Open Reading Frames (ORFs) during the cell replication cycle in the budding yeast (Saccharomyces cerevisiae). The cell cycle consists of four distinct phases: G_1 phase, DNA synthesis (S) phase, G_2 phase (also known as interphase) and mitosis (M) phase (Spellman et al., 1998). The microarray data set consists of 77 samples collected with three different synchronisation experimental conditions (alpha, cdc and elu). For the binding information, we utilised the ChIP-on-chip experimental result from (Lee et al., 2002), which provides significant levels (*p*-values) of 113 TFs binding to their target genes. We took negative of logarithm (base 10) of *p*-values to convert the significant levels to binding strengths. After mapping these two data sets, we finally obtained 6,099 ORFs that have expression measurements and binding information simultaneously. Among them, approximately 800 genes have been identified as cell cycle-regulated genes (Spellman et al., 1998). The goal of this study is to identify the cell cycle-related condition-specific gene modules, demonstrating the feasibility of the proposed method.

A breast cancer cell line microarray data set (Creighton et al., 2006) was further utilised in this study to identify condition-specific regulatory modules related to breast cancer. The original profiling study was designed to examine how estrogen-induced mRNA expression patterns observed in *in vitro* cell line models correlate with the expression patterns in breast tumours *in vivo*. Estrogen plays a significant role in breast cancer development and progression. The authors in Creighton et al. (2006) treated three estrogen-dependent breast cancer cell lines (MCF-7, T47D and BT-474) with 17β-estradiol (E2) and profiled the gene expression using Affymetrix Genechip Arrays. As reported in Creighton et al. (2006), eight E2-induced gene clusters were formed and their biological function was annotated by Gene Ontology (GO) terms. Among them, the expression patterns in two clusters (i.e., Cluster B and Cluster D as in Creighton et al. (2006)) clearly showed early up-regulation and late up-regulation, respectively. Significant GO terms in Cluster B are related to ‘ribosome biogenesis’, ‘RNA metabolism’ and ‘protein folding’, while significant GO terms in Cluster D include ‘cell cycle’, ‘cell proliferation’, ‘mitosis’ and ‘DNA replication’. After mapping the expression data with motif binding strength data, we finally obtained gene expression measurements with 39,407 probe sets and their corresponding binding strengths with 586 motifs. Specifically, the number of probe sets in early up-regulation stage is 692 and the number in late up-regulation stage is 334.

3.2 Identifying cell cycle-related regulatory modules in yeast

We used MATLAB SVM toolbox (Gunn, 1997) to implement the ε-insensitive linear SVR and SOM clustering algorithm (Kohonen, 1997) to form multi-level gene clusters. The parameters in the algorithm were empirically determined for this experiment. Specifically, we set the permutation parameter B (see equation (9)) as 1000, threshold t_0 (see Step 4 in the multi-level analysis procedure) as 50, and *p*-value threshold for each level as $P_0^1 = \min(1, 0.01 + 0.001 \times (L - l))$. Since the clustering method (i.e., SOM) generated slightly different results depending on its random initialisations, we repeated the whole procedure ten times with different initialisations in search for more reliable results. The significant motif sets and their regulatory modules were determined according to their average values of ten different initialisations. We also compared *k*-means (Hartigan and Wong, 1978) clustering results with SOM results after multiple initialisations and found the overlap rate is greater than 90% in terms of clustering membership, which would lead to the similar results for regulatory module identification. Therefore we only analysed the results based on SOM clustering method in follows.

We first calculated the significant level of each individual transcription factor on the cell cycle gene set using our approach. Table 1 lists the top 15 regulators, their functional descriptions, and their average *p*-values (<0.01). Among these regulators, 13 have been biologically validated as cell cycle-related regulators (Lee et al., 2002) and their phases of cell cycle are also shown in the table. The number of genes regulated by each regulator is also shown in the table. For those cell cycle-related regulators, the heatmap of their target gene expression profiles is shown in Figure 2, ordered by cell cycle phases in which the regulators are activated. From the heatmap we can clearly observe the cell cycle pattern of those targets genes across different phases, which demonstrated the effectiveness of our method for condition-specific regulatory module identification.

We then searched for cooperative TFs through a greedy forward search strategy. The upper limit number of regulators was set to 3 in this study. Table 2 lists the found cooperative TFs, their functional descriptions, the average *p*-values across all levels and the number of genes in the regulatory modules. In Table 2, several cooperative TFs have been demonstrated to be cell cycle-related in the previous studies (Tsai et al., 2005). For instance, DIG1-STE12, MBP1-SWI6, ACE2-SWI5 and SWI4-SWI6 (SBF) are known as synergistic pairs to regulate genes at cell cycle phases and they are also stably and significantly shown in our study. Interestingly, neither DIG1 nor STE12 is included in our individual regulator list (Table 1). Rather, the interaction between them showed significant regulation effect. More specifically, DIG1 is involved in the MAP kinase (MAPK) signalling pathway to regulate STE12, which is responsible for activating genes in response to MAP kinase cascades controlling mating and filamentous growth (Olson et al., 2000). MBP1 forms a complex (MBF) with SWI6 that binds to MluI cell cycle box regulatory element in the promoters of DNA synthesis genes. DNA binding component of the SBF complex (formed by SWI4 and SWI6) is a transcriptional activator that in concert with MBF regulates late G_1-specific transcription of targets including cyclins and genes required for DNA synthesis and repair (Ubersax et al., 2003). ACE2 and SWI5 bind the same DNA sequences *in vitro* with similar affinities, regulating a shared set of genes *in vivo* (McBride et al., 1999). In addition, MCM1, FKH1 and FKH2 are the critical activators of a group of M phase-specific transcripts (Breeden, 2003) and we found in this study that these three TFs formed a cooperative group to regulate their target genes.

3.3 Identifying condition-specific regulatory modules related to breast cancer

As a pre-processing step, we took the average of expression levels across all samples as the control value for each gene to calculate the log ratio data. The parameters in the algorithm were again empirically determined for this experiment as described in Section 3.2. We set the *p*-value threshold for each level as $p_0^l=\min(1, 0.05+0.01 \times (L-l))$. Similarly, we repeat the whole procedure ten times with different clustering initialisations for a more reliable result. The significant motif sets and their regulatory modules were selected according to their average values of the ten different initialisations.

Tables 3 and 4 list the identified motif sets for early and late stages, respectively. For each individual significant motif set, we list their motif identifiers, corresponding TFs, the average *p*-value across all levels and the number of genes in the regulatory modules. From Tables 3 and 4, we can see that the significantly enriched motif sets are quite different for early and late stages, which indicates the condition-specific nature of transcriptional regulation. Among them, we have found that many motif sets are biologically meaningful as reported in previous studies. For instance, STAT5A belongs to Jak2/Stat5 signalling pathways that plays a central role in principal cell fate decisions, regulating the processes of cell proliferation, differentiation and apoptosis (Wagner and Rui, 2008). E2F and Sp1/Sp3 have been shown to be synergistic to control gene expression (Kramps et al., 2004). Nuclear factor 1 (NF-1) family members interact with hepatocyte nuclear factor 1 (HNF-1) to synergistically active L-type pyruvate kinase gene transcription (Satoh et al., 2005).

For a more detailed analysis, we focus on the most significant single motif in each stage. In the early stage, the most significant module is regulated by c-Myc through binding site V $MYCMAX_03. Figure 3(a) shows several expression profiles associated with c-Myc, indicating the average gene expression profile of the module, the gene expression profile of c-Myc and the estimated transcription factor activity. From the figure, we can clearly see that the gene expression pattern of c-Myc is quite consistent with its estimated transcription factor activity, which is over-expressed in very early stage. c-Myc is a proto-oncogene that is amplified and plays a role in amplification of multiple other genes in breast cancer (Liao and Dickson, 2000). The significant functional annotations for this module include ‘nucleic acid binding’, ‘Ribosome’ and ‘RNA binding’ as obtained from the David database (Dennis et al., 2003). We further examined this gene module with Ingenuity Pathways Analysis and Figure 3(b) shows a c-Myc-involved network related to cancer, tumour morphology, cellular growth and proliferation.

Similarly, in the late stage, we found a module regulated by Oct-1 through the binding site V $OCT1_06, which was significantly enriched in multi-level clusters. Figure 4(a) shows several different expression profiles related to OCT-1. The Oct-1 gene expression profile is very much consistent with its estimated TFAs. Octamer transcription factor-1 (Oct-1) is a member of the POU family of TFs, and is involved in the transcriptional regulation of a variety of gene expressions related to cell cycle regulation, development, and hormonal signals (Kakizawa et al., 2001). The significant functional annotations from the David database for this module include ‘nucleic acid binding’, ‘cell cycle’ and ‘mitosis’. Figure 4(b) also shows a network extracted from Ingenuity Pathways Analysis for the OCT-1 module, which is largely related to gene expression, cancer and cell cycle.

4 Conclusions

Identification of transcription regulatory module has become increasingly important to understand the underlying mechanisms related to cancers. However, it is a quite challenging problem due to many noises in data sources and little knowledge available for data integration. In this paper, we have proposed a new method, namely multi-level regulatory module identification, to identify condition-specific gene regulatory modules. Motif binding information and gene expression profiles are integrated by SVR followed by significance analysis to find the active motif sets. A multi-level analysis strategy is further developed to help reduce false positives for reliable regulatory module identification. The method has been applied to a yeast cell cycle data set and a breast cancer microarray data set to identify the condition-specific regulatory modules. The experimental results show that our method can be reliably used to identify biologically meaningful regulatory modules. The regulatory modules identified from the breast cancer study may have important implications to understanding the pathways associated with breast cancer. In future work, the regulatory module identification method by SVR could be improved through iteratively updating regulatory binding strength and more simulation experiments are needed in order to compare with other methods.

Acknowledgments

This research was supported in part by NIH Grants (NS29525-13A, EB000830, CA109872 and CA096483) and a DoD/CDMRP grant (BC030280).

Biographies

Li Chen received her BA and MA in Computer Science and Engineering from Beijing Information Technology Institute and Shanghai Jiaotong University in 2000 and 2004, respectively She received MA in Computer Science and Engineering from University of South Florida in 2006. She is currently a PhD student in Department of Electrical and Computer Engineering in Virginia Polytechnic Institute and State University. Her areas of interest are machine learning, data mining, bioinformatics, computational biology and system biology.

Jianhua Xuan received his PhD Degree in Electrical Engineering and Computer Science from the University of Maryland in 1997. He received his BS, MS, and PhD Degrees from University of Zhejiang, China, in 1985, 1988, and 1991, respectively, all in Electrical Engineering. Currently, he is an Associate Professor of Electrical and Computer Engineering at Virginia Polytechnic Institute and State University. His research interests include biomedical image analysis, cellular and molecular imaging, computational bioinformatics, systems biology, intelligent information systems, visual intelligence, computer vision, information visualisation, and machine learning.

Yue Wang received his BS and MS Degrees in Electrical and Computer Engineering from Shanghai Jiao Tong University in 1984 and 1987, respectively. He received his PhD Degree in electrical engineering from University of Maryland Graduate School in 1995. Currently, he is a Professor of Electrical, Computer, and Biomedical Engineering at Virginia Polytechnic Institute and State University. His research interests focus on intelligent computing, machine learning, pattern recognition, statistical visualisation, and advanced imaging and image analysis, with applications to molecular analysis of human diseases.

Eric P. Hoffman received his PhD Degree in Biology/Genetics from Johns Hopkins University in 1986. He received BA Degrees in both Biology and Music from Gettysburg College in 1982. From 1986–1988 he was a Post-Doctoral fellow with Louis Kunkel at Harvard Medical School, and Boston Children’s Hospital. Since 1990, he has been Professor of Pediatrics at George Washington University School of Medicine and Health Sciences, and Director of the Research Center for Genetic Medicine at Children’s National Medical Center in Washington DC. His research interests include molecular pathogenesis of muscle disease, exercise physiology, development of novel therapeutics, and bioinformatics.

Rebecca B. Riggins received her BA in Biochemistry from Hood College (Frederick, MD USA) in 1998, followed by her PhD in Microbiology from the University of Virginia in 2003. From 2003 to 2006 she was a postdoctoral fellow in the Department of Oncology at Georgetown University, and in 2006 she was appointed Research Assistant Professor in this department. Her research is focused on understanding how lobular breast cancer responds (or becomes resistant) to endocrine therapies, combining laboratory research in breast cancer model systems with bioinformatics analysis of clinical breast cancer data.

Robert Clarke earned a DSc in 1999, a PhD in 1986, and a MSc in 1982 (each in Biochemistry) from the Queen’s University of Belfast (UK) and a BSc (Biological Sciences) in 1980 from the University of Ulster (UK). He completed his postdoctoral training at the Medical Breast Section of the National Cancer Institute as a Breast Cancer Study Group Fellow (1988). Currently he is a Professor of Oncology and Professor of Physiology and Biophysics at Georgetown University. He is working on the development and application of genomic and bioinformatic methods in translational studies in both humans and experimental models.

References

Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B. Computational detection of cis -regulatory modules. Bioinformatics. 2003; 19(Suppl. 2):ii5–ii14. [PubMed: 14534164]

Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 2006; 34:W369–W373. Web Server issue. [PubMed: 16845028]

Breeden LL. Periodic transcription: a cycle withm a cycle. Current Biology. 2003; 13(1):31–38.

Chekmenev DS, Haid C, Kel AE. P-match: transcription factor binding site search by combining patterns and weight matrices. Nucleic Acids Res. 2005; 33:W432–W437. Web Server issue. [PubMed: 15980505]

Creighton CJ, Cordero KE, Larios JM, Miller RS, Johnson MD, Chinnaiyan AM, Lippman ME, Rae JM. Genes regulated by estrogen in breast tumor cells in vitro are similarly regulated in vivo in tumor xenografts and human breast tumors. Genome Biol. 2006; 7(4):R28. [PubMed: 16606439]

Das D, Nahle Z, Zhang MQ. Adaptively inferring human transcriptional subnetworks. Mol Syst Biol. 2006; 2:0029. [PubMed: 16760900]

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation Visualization, and Integrated Discovery. Genome Biol. 2003; 4:5, p3.

Drucker, H.; Burges, CJC.; Kaufman, L.; Smola, A.; Vapnik, V. Advances in Neural Information Processing Systems. Vol. 9. The MIT Press; 1997. Support vector regression machines; p. 155

Gunn, SR. Support Vector Machines for Classifications and Regression. Image Speech and Intelligent Systems Research Group, University of Southampton; 1997.

Hartigan JA, Wong MA. A K-means clustering algorithm. App Statist. 1978; 28:100–108.

Ihmels J, Bergmann S, Barkai N. Defining transcription modules using large-scale gene expression data. Bioinformatics. 2004; 20(13):1993–2003. [PubMed: 15044247]

Kakizawa T, Miyamoto T, Ichikawa K, Takeda T, Suzuki S, Mori J, Kumagai M, Yamashita K, Hashizume K. Silencing mediator for retinoid and thyroid hormone receptors interacts with octamer transcription factor-1 and acts as a transcriptional repressor. J Biol Chem. 2001; 276(13): 9720–9725. [PubMed: 11134019]

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ. The UCSC genome browser database. Nucleic Acids Res. 2003; 31(1):51–54. [PubMed: 12519945]

Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. 2003; 31(13): 3576–3579. [PubMed: 12824369]

Kohonen, T. Self-Organizing Maps. Springer; NY, NY: 1997.

Kramps C, Strieder V, Sapetschnig A, Suske G, Lutz W. E2F and Sp1/Sp3 Synergize but are not sufficient to activate the MYCN gene in neuroblastomas. J Biol Chem. 2004; 279(7):5110–5117. [PubMed: 14645238]

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. Transcriptional regulatory networks in saccharomyces cerevisiae. Science. 2002; 298(5594):799–804. [PubMed: 12399584]

Liao DJ, Dickson RB. c-Myc in breast cancer. Endocr Relat Cancer. 2000; 7(3):143–164. [PubMed: 11021963]

Lomax, RG. Statistical Concepts: A Second Course. Lawerence Erlbaum Associates; Mahwah, NJ: 2007.

Matys V, Kel-Margoulis OV, Fncke E, Liebich I, Land S, Barre-Dirne A, Reuter L, Chekmenev D, Krull M, Hormscher K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006; 34:D108–D110. Database issue. [PubMed: 16381825]

McBride HJ, Yu Y, Stillman DJ. Distinct regions of the Swi5 and Ace2 transcription factors are required for specific gene activation. J Biol Chem. 1999; 274(30):21029–21036. [PubMed: 10409653]

Olson KA, Nelson C, Tai G, Hung W, Yong C, Astell C, Sadowski I. Two regulators of Ste12p inhibit pheromone-responsive transcnption by separate mechanisms. Mol Cell Biol. 2000; 20(12):4199–4209. [PubMed: 10825185]

Qi Y, Ge H. Modularity and dynamics of cellular networks. PLoS Comput Biol. 2006; 2(12):e174. [PubMed: 17196032]

Ruan J, Zhang W. A bi-dimensional regression tree approach to the modeling of gene expression regulation. Bioinformatics. 2006; 22(3):332–340. [PubMed: 16303796]

Satoh S, Noaki T, Ishigure T, Osada S, Imagawa M, Miura N, Yamada K, Noguchi T. Nuclear factor 1 family members interact with hepatocyte nuclear factor 1alpha to synergistically activate L-type pyruvate kinase gene transcription. J Biol Chem. 2005; 280(48):39827–39834. [PubMed: 16204235]

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature. 2008; 451(7178):535–540. [PubMed: 18172436]

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. 2003; 34(2):166–176. [PubMed: 12740579]

Sharan R, Ovcharenko I, Ben-Hur A, Karp RM. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. Bioinformatics. 2003; 19(Suppl. 1):i283–i291. [PubMed: 12855471]

Smola, AJ.; Scholkopf, B. NeuroCOLT2 Technical Report, Statistics and Computing. 1998. A Tutorial on Support Vector Regression.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 1998; 9(12):3273–3297. [PubMed: 9843569]

Tsai HK, Lu HH, Li WH. Statistical methods for identifying yeast cell cycle transcription factors. Proc Natl Acad Sci USA. 2005; 102(38):13532–13537. [PubMed: 16157877]

Ubersax JA, Woodbury EL, Quang PN, Paraz M, Blethrow JD, Shah K, Shokat KM, Morgan DO. Targets of the cyclin-dependent kinase Cdk1. Nature. 2003; 425(6960):859–864. [PubMed: 14574415]

Wagner KU, Rui H. Jak2/Stat5 signaling in mammogenesis, breast cancer initiation and progression. J Mammary Gland Biol Neoplasia. 2008; 13(1):93–103. [PubMed: 18228120]

Wang W, Cherry JM, Nochomovitz Y, Jolly E, Botstem D, Li H. Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. Proc Natl Acad Sci, USA. 2005; 102(6):1998–2003. [PubMed: 15684073]

Zhou Q, Wong WH. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc Natl Acad Sci, USA. 2004; 101(33):12114–12119. [PubMed: 15297614]

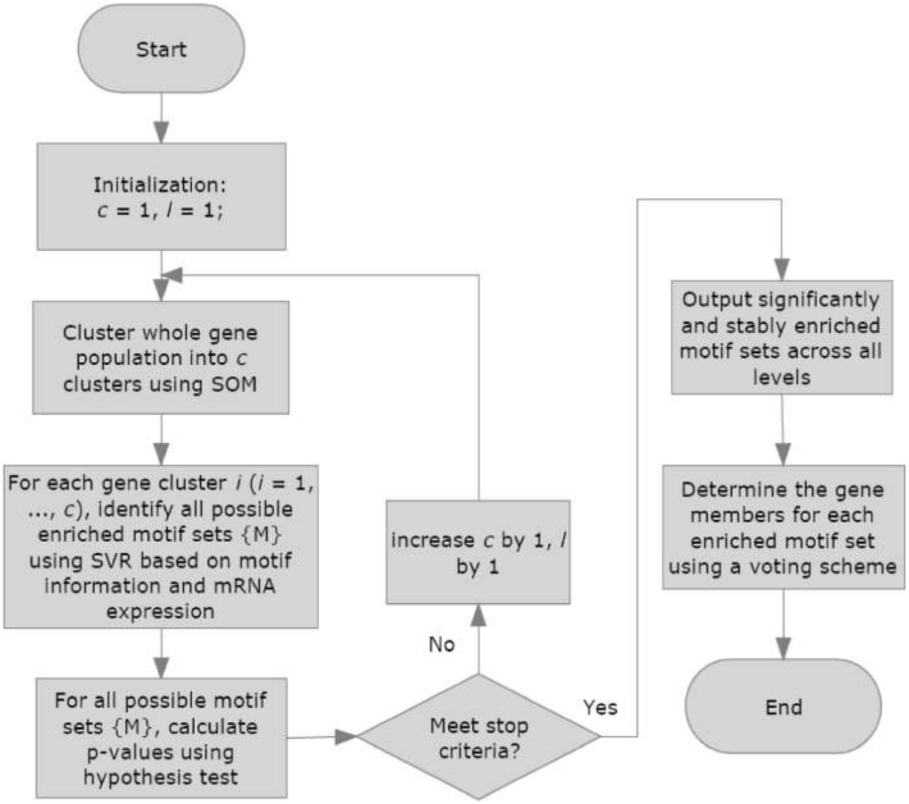

Start
Initialization:
c = 1, l = 1;
Cluster whole gene population into c clusters using SOM
For each gene cluster i (i = 1, ..., c), identify all possible enriched motif sets {M} using SVR based on motif information and mRNA expression
For all possible motif sets {M}, calculate p-values using hypothesis test
Meet stop criteria?
No
increase c by 1, l by 1
Yes
Output significantly and stably enriched motif sets across all levels
Determine the gene members for each enriched motif set using a voting scheme
End

Figure 1.

Flowchart of multi-level regulatory module identification procedure

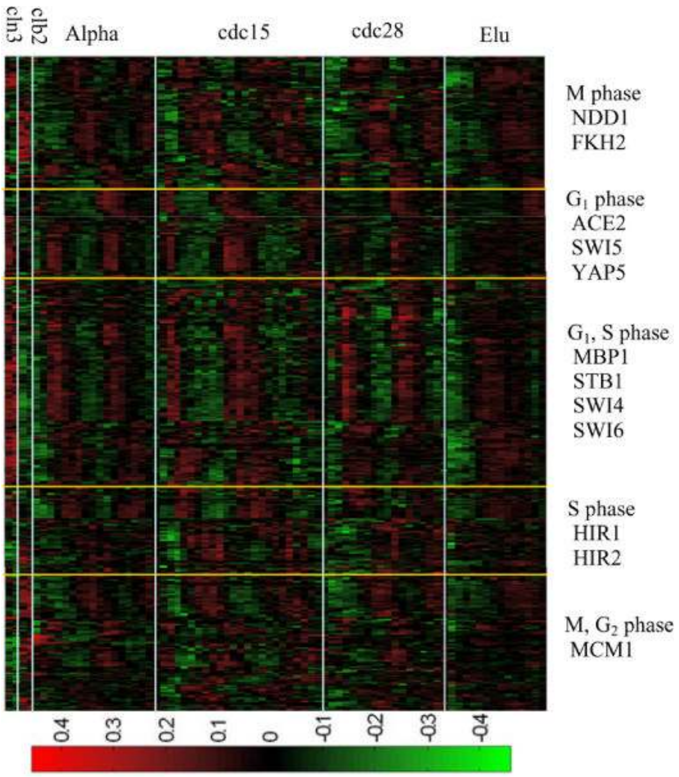
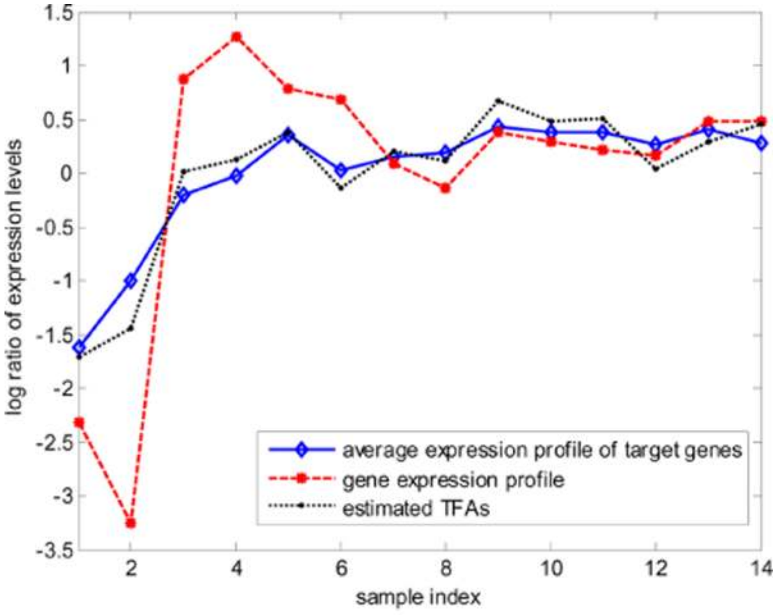


Figure 2.

Heatmap of gene expression profiles for regulatory modules at different cell cycle phases. Each row represents gene expression profile across different conditions and each column represents one sample. Genes are ordered by the cell cycle phased in which their regulators are activated (see online version for colours)


1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
log ratio of expression levels
2
4
6
8
10
12
14
sample index
average expression profile of target genes
gene expression profile
estimated TFAs

(a)

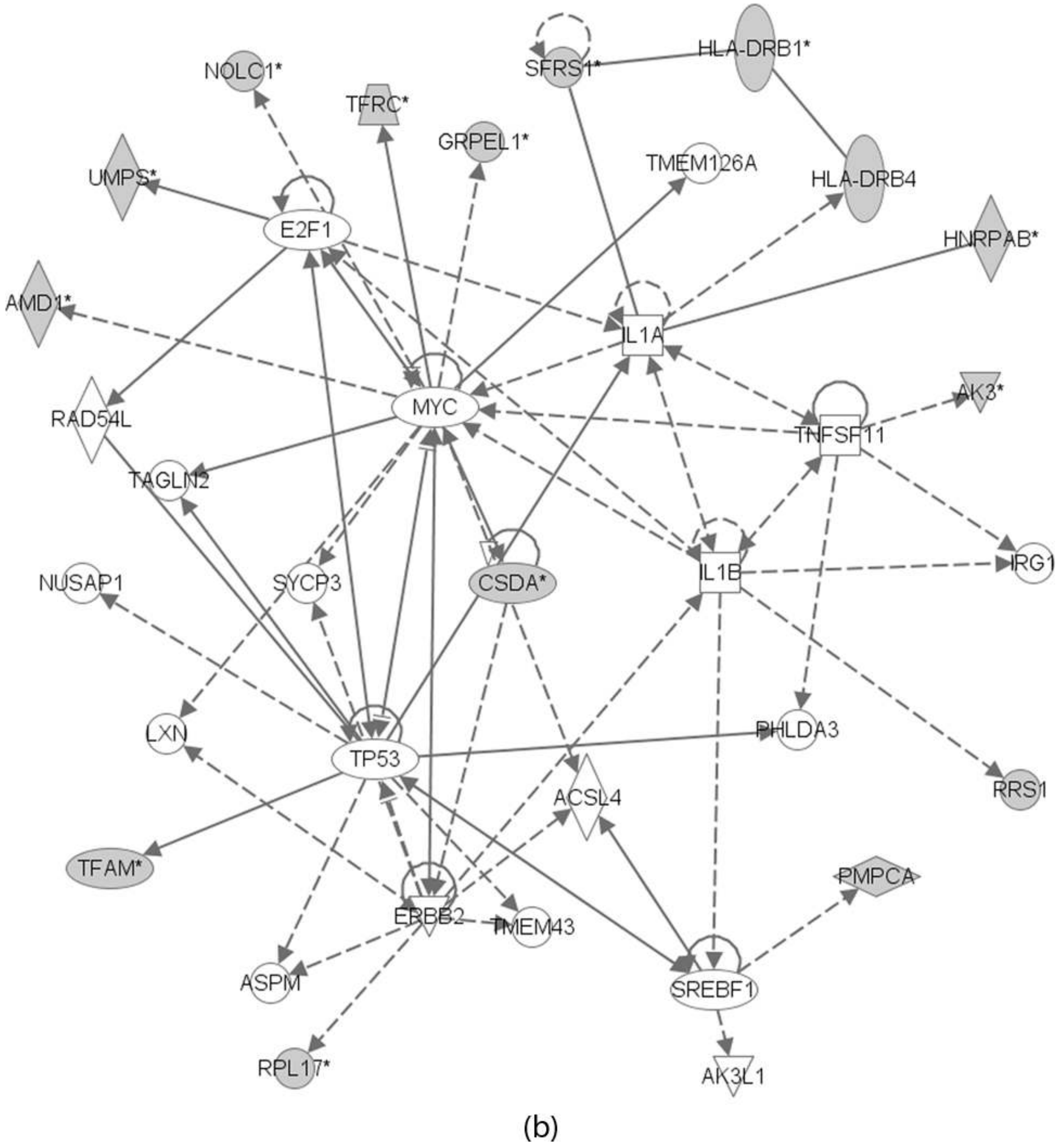
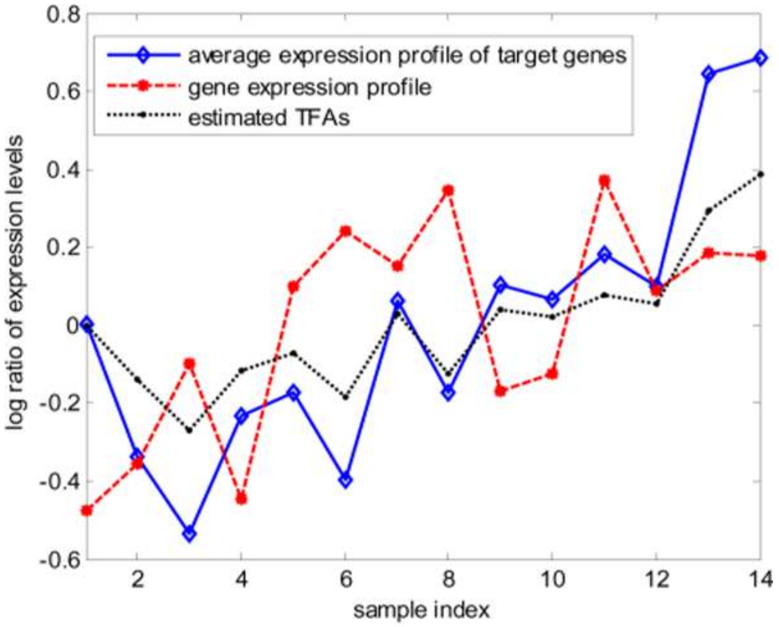


Figure 3.

The identified c-Myc module in the early stage: (a) the expression profiles indicating the average gene expression profile of the module, the gene expression profile of c-Myc and the estimated transcription factor activity and (b) the c-Myc network related to cancer, tumour morphology, cellular growth and proliferation (see online version for colours)


average expression profile of target genes
gene expression profile
estimated TFAs
log ratio of expression levels
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
2
4
6
8
10
12
14
sample index

(a)

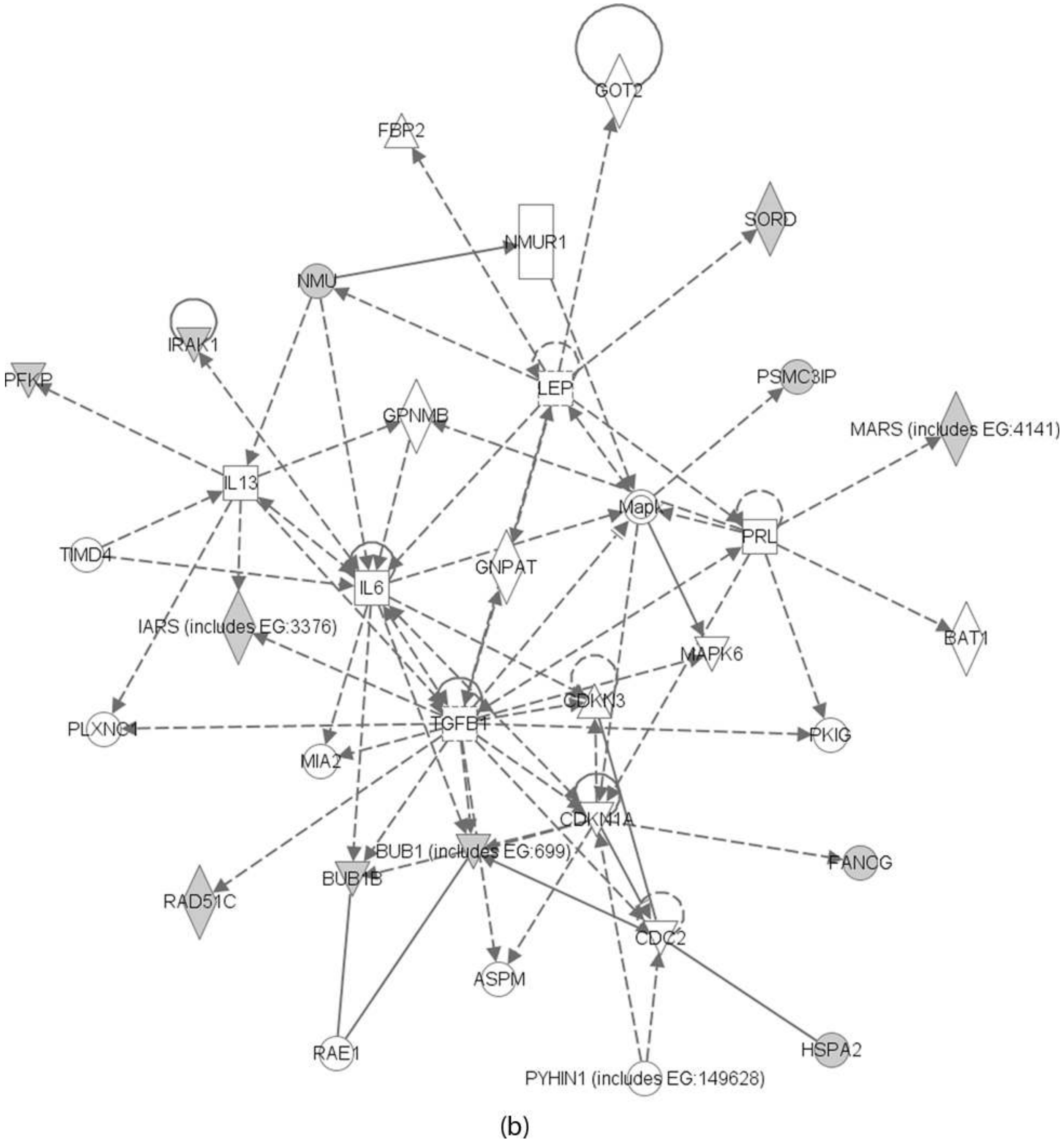


Figure 4.

The identified OCT-1 module in the late stage: (a) the expression profiles indicating the average gene expression profile of the module, the gene expression profile of OCT-1 and the estimated transcription factor activity and (b) the OCT-1 network related to gene expression, cancer and cell cycle (see online version for colours)

Table 1

Identified significantly enriched individual regulators in yeast cell cycle data set

TF	*Description*	*Phases*	*p-value*	*No. of genes*
NDD1	Activation of its M phase-specific target genes	M	0.0003	45
MCM1	Activator of G2 and M phase-specific transcripts	G_2, M	0.0004	72
FKH2	Activation of its M phase-specific target genes	M	0.0004	75
HIR1	Subunit of the HIR complex, a nucleosome assembly complex involved in regulation of histone gene transcription	S	0.0011	43
MBP1	Transcription factor involved in the regulation of cell cycle progression from G1 to S phase	G_1, S	0.0014	83
STB1	Protein with a role in regulation of MBF-specific transcription at Start, phosphorylated by Cln-Cdc28p kinases in vitro; unphosphorylated form binds Swi6p and binding is required for Stb1p function; expression is cell-cycle regulated	G_1, S	0.0017	79
SWI4	Involved in cell cycle-dependent gene expression	G_1, S	0.0023	79
HIR2	Subunit of the HIR complex, a nucleosome assembly complex involved in regulation of histone gene transcription; recruits Swi-Snf complexes to histone gene promoters	S	0.0027	53
ACE2	Activates expression of early G1-specifci genes	G_1	0.0029	43
SWI6	Forms complexes with DNA-binding proteins Swi4p and Mbp1p to regulate transcription at the G1/S transition; involved in meiotic gene expression	G_1, S	0.0036	80
SWI5	Activates expression of early G1-specific genes	G_1	0.0058	48
YAP5	bZIP transcription factor	G_1	0.0063	47
GAT3	Protein containing GATA family zinc finger motifs	–	0.0063	46
MET4	Lecine-zipper transcriptional activator, responsible for the regulation of the sulphur amino acid pathway	–	0.0066	88
GAL4	DNA-binding transcription factor required for the activation of the GAL genes in response to galactose	–	0.0072	58

Thirteen TFs (in boldface) have biological support to be cell cycle-related.

Table 2

Identified significantly enriched cooperative regulators in yeast cell cycle data set

Doublets

Transcription factor	*P-value*	*No. of genes*	*Transcription factor*	*P-value*	*No. of genes*
RGM1	0.0020	14	**MBP1**	0.0052	76
CBF1			**SWI6**		
DIG1	0.0032	17	**ACE2**	0.0076	46
STE12			**SWI5**		
MET4	0.0038	35	CRZ1	0.0093	32
YAP5			IXR1		
CHA4	0.0042	67	**SWI4**	0.0107	76
HAP4			**SWI6**		

Triplets

Transcription factor	*P-value*	*No. of genes*
HIR1	0.0016	91
YAP5		
HIR2		
FKH1	0.0024	107
MCM1		
FKH2		
PHO2	0.0057	68
SFL1		
RTG1		

Five regulator sets (in boldface) have literature support for their physical interactions or synergistic effects.

Table 3

Identified significantly enriched motif set in the early stage

Motif identifier	*Transcription factor*	*Average p-value*	*No. of genes in module*
Single			
V$MYCMAX_03	c-Myc	0.034	128
V$AHRARNT_02	AhR	0.040	60
V$PAX9_B	PAX	0.046	98
V$STAT5A_02	STAT5A	0.058	83
Doublets			
V$OSF2_Q6	AML3	0.032	52
V$AP2GAMMA_01	AP-2gamma		
V$E2F1_Q6_01	E2F	0.045	132
V$SP3_Q3	SP3		
V$STAT5A_04	STAT5A	0.074	138
V$ZF5_01	ZF5		
V$AP1_01	AP1	0.081	82
V$SOX5_01	SOX5		
Triplets			
V$TATA_01	TATA	0.040	58
V$IRF_Q6	IRF		
V$TBX5_Q5	TBX5		
V$STAT5A_01	STATA5	0.074	79
V$EGR3_01	Egr3		
V$XFD1_01	XFD1		
V$FOXO1_02	FOX	0.085	187
V$SP3_Q3	SP3		
V$NRF2_01	NRF-2		

Table 4

Identified significantly enriched motif set in the late stage

Motif identifier	*Transcription factor*	*Average p-value*	*No. of genes in module*
Single			
V$OCT1_06	Oct-1	0.055	60
V$NKX25_02	Nkx2-5	0.060	68
V$IK2_01	Ik-2	0.068	87
Doublets			
V$STAT5A_03	STATA5	0.0284	106
V$ZIC1_01	ZIC1		
V$HNF1_01	HNF-1alpha-A	0.100	12
V$NF1_Q6	NF-1		