

A novel method for studying transcription factor combinations as cell cycle-related modules from yeast genomic datasets

Journal:	*Nucleic Acids Research*
Manuscript ID:	NAR-00808-S-2007
Manuscript Type:	1 Standard Manuscript - US Editorial Office
Key Words:	Transcription Factors, Modules, Chromatin Immunoprecipitation

scholarONE™
Manuscript Central

For Peer Review

A novel method for studying transcription factor combinations as cell cycle-related modules from yeast genomic datasets

Jhih-Long Huang[1], Chea-Yuan Lee[1], Yi Huang[2], Hung-Mao Li[1] and Kuang-Den Chen[1,2*]

[1] Chang Gung Molecular Medicine Research Center, Chang Gung University, Taoyuan, Taiwan

[2] Department of Life Science, Chang Gung University, Taoyuan, Taiwan

To whom correspondence should be addressed to Dr. K.-D. Chen at Chang Gung Molecular Medicine Research Center, and Department of Life Science, Chang Gung University, 259 Wen-Hwa 1st Road, Kwei-shan, Taoyuan, Taiwan 333 and E-mail: ktchen@mail.cgu.edu.tw

For Peer Review

Abstract

Transcription regulation is known to occur through the cooperative and combined efforts of multiple transcription factors (TFs). Although previous studies have attempted to find cooperative TFs based on various genomic datasets, each dataset has provided only weak evidence, prompting researchers to seek methods for integrating various datasets for the identification of cooperative TFs. Here, we introduce a novel method for integrating genome-wide gene expression data and chromatin immunoprecipitation (ChIP) data to discover more biologically relevant combinations of TFs. We applied Apriori-like approach to integrate TFs with other TFs or with cell cycle phases as module instead of pairs, and then defined a novel measure of correlation among them to extract modules with biological significance. Using this technique, we identified some modules of TFs that appear to cooperatively regulate transcription in the yeast cell cycle, and assigned them to the appropriate phase of the cell cycle. We then used these modules to construct a new combinatorial regulatory map of the yeast cell cycle. Our results also identified modules consistent with previous reports of factors that play roles in biological processes such as metabolism and the stress response. Thus, our novel method successfully identified modules in good agreement with existing reports, as well as previously unidentified modules that may prompt new avenues of study.

For Peer Review

Introduction

In higher eukaryotes, transcriptional regulation of genes is controlled by multiple transcription factors (TFs) and the interactions among them. TFs may be divided into two categories based on their binding interactions, namely DNA-binding factors, which separately bind their own motifs on the same promoter, and non-DNA-binding factors, which do not directly interact with DNA, but instead interact with a DNA-binding factor capable of binding DNA. There are many types of interactions between TFs and target genes, each TF may regulate many target genes, and each target gene may be controlled by multiple TFs. Thus, a full understanding of transcriptional regulatory networks requires knowledge of combinatorial interactions among many TFs.

Researchers have developed a number of methods for inferring cooperativity among TFs (1-7). One recently developed genome-based technique is the use of DNA microarray analysis to identify TF binding motifs by primary clustering genes according to similar expression patterns (8,9), and then using motif-finding algorithms based on gene promoter sequences to identify motifs that are over-represented in a given cluster, which are then taken to be TF binding motifs (10-14). However, this approach is limited in that it does not allow identification of cooperative combinations of multiple TFs. Another recently developed genome-based technique, chromatin immunoprecipitation (ChIP), does allow a large number of TFs to be connected to a large number of target genes. Lee *et al.* (15) published ChIP microarray data for most of the yeast TFs, and identified 113 yeast cell cycle-related TFs. The ChIP microarray technique is thought to provide reliable evidence of functional TF-binding site interactions and identify direct binding of a specific protein complex to DNA. Furthermore, although ChIP datasets and expression datasets are different kinds of data obtained from different techniques, it is possible to combine these datasets to predict

For Peer Review

cooperativity among TFs (1,2,5). Proper integration of these datasets can allow researchers to not only identify cooperative TFs, but also to assign them to appropriate biological processes and obtain far greater statistical power than that generated by non-integrated datasets.

In this paper, we used Apriori-like approach to integrate a ChIP dataset with a gene expression dataset, and predicted cooperative combinations among multiple TFs associated with the yeast cell cycle. In our analysis, a combination of TFs was considered to be cooperative if the intersection of their target genes had statistical correlation, which we assessed with a novel measure. Additionally, we also examined statistical correlations among the intersections between the combined TFs and some cell cycle phase-specific target genes. A high correlation was taken to indicate that the combination of TFs was involved in the given cell cycle phase. Thus, we were able to not only identify cooperative TFs, but also connect them with the corresponding cell cycle phase as a module. Application of this method to ChIP and expression microarray data pertaining to yeast cell cycle genes allowed us to identify some modules of TFs cooperating to regulate yeast cell cycle-associated transcription, and further allowed us to assign them to the appropriate cell cycle phase. Many of the identified modules were consistent with reports in the literature, validating our approach. In addition, we identified several novel TF modules that appear to be directly or indirectly related to the cell cycle, offering new associations that may be tested in future studies. Finally, we were able to combine all of the cell cycle-related modules to propose an integrated transcriptional regulation model of the yeast cell cycle. Thus, we herein report a novel and integrated approach for combining results from diverse genome-wide datasets, and show that this integrated approach can substantially eliminate noises contained in each individual dataset and improve the accuracy when modeling.

For Peer Review

Materials and Methods

The standard regulatory model of the yeast cell cycle

According to the traditional regulatory model of transcriptional regulators in the yeast cell cycle, MBF (a complex of MBP1 and SWI6) and SBF (a complex of SWI4 and SWI6) control the late G1 genes. Mcm1, together with FKH1 or FKH2, recruits the NDD1 protein in late G2, and thus controls transcription of the G2/M genes. MCM1 is also involved in the transcription of some M/G1 genes, while SWI5 and ACE2 regulate genes at the end of M and early G1 (17). Some studies have suggested that the SBF and FKH1/FKH2 protein complexes regulate gene expressed during the S/G2 phase (15, 17, 18). STB1 binds to SWI6 and is indirectly thought to be involved in the G1 phase (19), and HIR1 and HIR2, which act as transcriptional repressors, regulate histone gene transcription during cell cycle and trigger histone synthesis in the S phase (20).

Data preprocessing

We used the cell cycle gene expression data published by Spellman *et al*. (16), in which 800 yeast cell cycle genes were identified and divided into subsets relevant to the G1, S, S/G2, G2/M and M/G1 phases of the cell cycle. We also used a genome-wide ChIP dataset containing information on 113 TFs, as reported by Lee *et al.*(15). We identified and defined as TF-targeted genes for each of the 113 TFs in the ChIP dataset using provided binding probability with a threshold of P_b <0.01, which was less stringent than that recommended in the previous report (P_b <0.001) (15). We regarded the TFs and cell cycle phases as “items”, which were then used to form modules in our study.

For Peer Review

Assessing correlation among items

We assessed correlations among items using the following statistical method. For the purpose of this analysis, a gene was said to occur on an item if it was bound by this item (in the case of a TF) or had been previously classified with the item (in the case of a cell cycle phase). A pair of items was considered correlative if a set of genes related to both items was significantly greater than the set of genes related to either item alone. The probability (*P*-value) of observing at least m genes from the set of genes related to the second item within a set of n genes related to the first item in hypergeometric distribution and is given by:

$$p-value = 1 - \sum_{i=0}^{m-1} \frac{\binom{f}{i} x \binom{N-f}{n-i}}{\binom{N}{n}} \qquad (1)$$

Where f is the total number of genes in the set belonging to the second item and N is the total number of genes related to at least one of the items. Thus, we calculated the probability that a given gene (m) would be related to both items by chance. In this case, it did not matter which item of the pair was regarded as the first item; the *P*-values were almost equal when the calculations were run with either item regarded as the first item (data not shown).

Instead of generating association rules among items, we then used an a-priori-like algorithm (17) to form large correlation modules containing the items (TFs and phases). We regarded TFs and phases as items and genes as transactions. If a given gene was bound by an item (TF) or defined as being classified with an item (phase), we said that the item (TF or phase) was bought with the transaction (gene). An a-priori algorithm typically seeks to identify modules containing items that are frequently bought together, but we expanded upon this by further asking whether the co-occurrence of a given

For Peer Review

group of items has any significance. To form the largest correlative modules using our Apriori-like algorithm, we first identified intersections between TF-targeted genes or between a TF-targeted gene and a phase-specific gene. We then measured the correlations of these intersections, and combined highly correlative intersections to yield 2-order modules, as statistically calculated using equation (1), where m is the number of genes total in a given pair of intersections. Next, we merged associated 2-order modules into extended 3-order modules (for example merging [I_1, I_2] and [I_2, I_3] into [I_1, I_2, I_3]), and similarly measured the correlation of the intersections in the extended 3-order modules. In this case, the number n of interacting genes within a given 3-order module was regarded as m for equation (1), and we calculated pairwise correlations between items according to equation (1), using n. Since $m \geq n$, the original pairwise correlation calculated using m was always less than that calculated using n. We then compared the pairwise correlations of each 3-order module to the threshold P_c. If the correlation was greater than the threshold value, then the 3-order module failed to form, despite the correlations among all of its 2-order modules. If the correlation was below the threshold, we regarded the extended 3-order module as a correlative 3-order module. Fig. 1 show a schematic depiction of the process of merging 2-order associated modules into 3-order extended modules and assessing their correlation. We then took each correlative 3-order extended module and attempted to merge it with other correlative 3-order modules in an effort to extend each module further. In cases where larger correlative modules failed to form with correlations below the threshold value, we defined the module as a largest correlative module. For example, modules, [I_1, I_2, I_3], [I_1, I_2, I_4], and [I_1, I_2, I_5] could theoretically be merged into extended modules [I_1, I_2, I_3, I_4], [I_1, I_2, I_3, I_5] and [I_1, I_2, I_4, I_5]. However, if [I_1, I_2, I_3, I_4] and [I_1, I_2, I_3, I_5] fail to pass the threshold test, then [I_1, I_2, I_3] must have failed to

For Peer Review

form any larger correlative modules, and may thus be extracted as a largest correlative module. This merging process was repeated until no more extended modules could form, meaning that the remaining non-extracted modules could be regard as the largest correlative modules. In the above case, [I_1, I_2, I_4, I_5] would be considered a largest correlative module, because it can not be extended further.

The largest correlative modules are the final desired results, wherein the largest possible numbers of correlative items are grouped together and no more grouping is possible. To obtain the largest correlative modules, it is necessary to search all possible modules of items. However, given t items, the number of candidate modules is $\left(\frac{t}{1}\right)+\left(\frac{t}{2}\right)+\left(\frac{t}{3}\right)+.....+\left(\frac{t}{t}\right)$. Hence, this strategy requires efficient algorithms capable of restricting the search space and checking only a subset of modules without missing important largest correlative modules. Instead of creating association rules among items, we used an Apriori-like algorithm (see Algorithm 1) to find the largest correlative modules. There are two differences between our Apriori-like algorithm and the original algorithm. First, while the original Apriori algorithm used a pruning principle to remove any extended modules with less than the minimal required support, we tested items among an extended module for pairwise correlations calculated by equation (1) and better than the threshold P_c. However, we used the original pruning principle to remove any extended modules containing non-correlative sub-modules. For example, if correlative 2-order modules [I_1, I_2] and [I_2, I_3] were merged into [I_1, I_2, I_3], but sub-module [I_1, I_3] was not correlative, then we would prune [I_1, I_2, I_3]. The second difference between our method and the original Apriori was that the original one was aimed at generating rules among found frequency item-sets, whereas our strategy was aimed at finding the largest correlation modules. For this, we defined our

For Peer Review

largest correlation modules as follows:

Definition 1. *A largest correlation module is one in which the maximum number of correlative items has been grouped together as a module, and no other items may be grouped into this module at the given threshold value.*

In particular, we were interested in finding phase-specific modules and TFs satisfying the following definitions:

Definition 2. *A phase-specific module is a largest correlative module containing any phase items. A phase-specific TF is a TF item that combines with a phase item as a largest correlative module.*

However, some of the largest correlative modules identified in the present work were not directly related to the yeast cell cycle, but could be indirectly associated with some phase of cell cycle, i.e. through protein-protein interactions. Therefore, we also sought phase-associate modules satisfying the following definition:

Definition 3. *A phase-associate module was the largest correlative module that lacked a phase item but contained a phase-specific TF.*

For Peer Review

Results

Preprocessing yeast CHIP and microarray data and assessing correlation probabilities

Primarily, the yeast genome-wide ChIP location dataset from Lee *et al.* (15) was used to identify cooperativity and all possible TF modules among the 113 listed TFs. The yeast cell cycle gene expression dataset from Spellman *et al*. (16) was further integrated to find the corresponding cell cycle phase of any correlative TF modules, thus generating a second set of correlative modules (see Materials and Methods). Among 6272 non-redundant genes in the Lee *et al*. (15) dataset, we identified 4421 TF-targeted genes that bound at least one of the 113 TFs in the ChIP dataset with *P*-value smaller than a threshold value (e.g., P_b <0.01). In the dataset of Spellman *et al*. (16), 800 identified cell cycle genes had been divided into five cell cycle phase-related subsets, namely G1, S, S/G2, G2/M and M/G1. We used provided probability threshold for TF binding to identify the intersection of the TF-target gene sets with the phase-specific genes, and used calculated probability threshold for interactive association to identify intersections with high correlations. We then used Apriori-like algorithm to combine TFs with other TFs (within the first dataset) and with specific phases (integrating the datasets), thus generating modules (see Materials and Methods). The *P*-value threshold (P_b) for TF binding to chromatin and the correlation *P*-value threshold (P_c) for forming modules were both defined as <0.01. The number of total genes for P_c calculation was 4577, which was the number of non-redundant genes bound by at least 1 TF (4421 genes) plus the number of phase-specific genes (800 genes). The other filtering criterion ($I_{\min}$) reflects whether or not a given module contains at least five genes ($I_{\min} = 5$). Using these criteria, we obtained 909 modules with correlations. Of them, 43 were phase-specific or phase-associate modules (

For Peer Review

Table 1). The identified subset of 909 modules is much smaller than the overall $[\binom{118}{1}+\binom{118}{2}+\binom{118}{3}+.....+\binom{118}{118}]$ possible modules that could be derived from 113 TFs and 5 cell cycle phases. It also provides a reasonably small set of modules that may be tested experimentally.

Phase-specific and phase-associate modules related to the yeast cell cycle

To determine the TF-targeted genes in association with specific phase, we organized the analyzed data in phase-specific and phase-associated manners in series as shown in Table 1 and Supplementary Table 1. In yeast G1 phase, MBF is known as composed of MBP1 and SWI6, while SBF is composed of SWI4 and SWI6; both MBF and SBF play a part in the G1 phase (18,19). STB1 binds to SWI6 and is thought to interact with SBF and MBF to regulate transcription in G1 phase (20,21). The analyses identified 8 modules in association with G1 phase, where 5 of these modules, [G1, STB1], [G1, MBP1, SWI6], [G1, SWI4, SWI6], [STB1, SWI6] and [MBP1, STB1, SWI4], were in good agreement with the standard regulatory model of the G1 phase. Among them, [STB1, SWI6] and [MBP1, STB1, SWI4] were phase-associate modules, whereas STB1, SWI4, SWI6 and MBP1 were phase-specific TFs that formed modules with one other (see Definition 3 in Materials and Methods). Although modules [G1, STB1] and [STB1, SWI6] did not merge into [G1, STB1, SWI6], which would have directly indicated that STB1 and SWI6 form a complex involved in the G1 phase, we could indirectly deduce such an interaction by observation of the phase-specific and phase-associate modules. In contrast to the above well-supported modules, two other identified modules, [G1, MBP1, SWI4] and [MBP1, SWI4, SWI6], were somewhat inconsistent with previous reports. These modules also appear to indirectly interact

For Peer Review

suggesting possibility of novel interactions. In addition, we also found a novel module [G1, YAP5] that is consistent with the report of Horak *et al.* (22) suggesting that YAP5 may participate in the more general processes of metabolism and respiration, and perhaps even cell growth during the G1/S transition.

In the S phase, HIR1/HIR2 are transcriptional co-repressors responsible for regulating histone gene transcription (23). Transcriptional regulation of the sulfur amino acid pathway in yeast depends on the activator, MET4, whose function requires different combinations of the factors CBF1, MET28, MET31 and MET32 (24). Two different molecular mechanisms may be used to tether the MET4 transcriptional activator to DNA, depending on the promoter region: one acting through CBF1/MET28 and the second requiring either MET31 or MET32 together with MET28 (24). MET28 and MET32 were not among the 113 TFs assayed by ChIP (15), and thus were not included in the present analysis. In the analyses we generated 8 modules in association with S phase. The phase-specific module, [S, HIR1, HIR2], and the phase-associate modules, [MET4, CBF1] and [MET4, MET31], were consistent with the standard regulatory model. In addition, our analysis newly predicted the involvement of MET4, MET31 and CBF1 in the S phase. Although this association has not been previously reported, the MET genes and histone genes are two major class of S phase-specific genes, and previous studies have shown that MET4, MET31 and CBF1 are associated with the transcriptional induction of the MET genes in response to low levels of adenosyl methionine (25,26). Thus, it seems logical that MET4, MET31 and CBF1 could be regulators in the S phase, as indicated by our analysis. We also identified a novel S phase-specific module, [S, MET4, HIR2]. Although no previous study has examined the involvement of MET4 in histone gene regulation, the binding motifs for MET4 exist in histone promoters, suggesting that MET4 may be a novel regulator of

For Peer Review

histone expression (1) in the S phase.

When we examined genes specific to the S/G2 phase we identified 3 modules in association with S/G2. Among them, several contained the forkhead genes, FKH1 and FKH2, which are associated with expressed TF-targeted genes during this cell cycle phase (27), wherein they bind forkhead sites with the same affinity (28). Thus, the phase-specific modules, [S/G2, FKH1] and [S/G2, FKH2], and phase-associate module [FKH1, FKH2], were consistent with the evidence found in the literature, as well as with the standard regulatory model of yeast cell cycle genetics.

Examination of genes involved in the G2/M phase yielded 8 G2/M-specific modules and 10 G2/M-associated modules. One identified module, [G2/M, MCM1, NDD1, FKH2], was consistent with experimental evidence that a complex containing MCM1, NDD1 and FKH1/FKH2 controls the transcription of G2/M genes (27,29). In contrast to the other cell cycle phases, however, the remaining identified models lacked support within the literature (see Supporting Materials), at least in part due to the general lack of published information regarding transcriptional control in the yeast G2/M phase (3).

When we examined genes involved in the M/G1 phase we identified 5 M/G1-specific modules and 1 M/G1-associated module. Previous studies have shown that ACE2 and SWI5 cooperatively induce the expression of some M/G1 genes (27,30), while YOX1 and YHP1 bind to MCM1 and cooperatively control the transcription of several M/G1 genes (27,31). Although YOX1 and YHP1 were not among the 113 TFs used herein (15), we identified modules [M/G1, MCM1], [M/G1, SWI5] and [M/G1, ACE2], which are in agreement with the standard regulatory model of the M/G1 phase. Module [M/G1, SWI5, ACE2] would have been even more consistent with previous reports; we did not observe this module using a correlation threshold of $P_c < 0.01$, but

For Peer Review

the use of the less stringent correlation threshold ($P_c \leq 0.1$) yielded this module (see below; Table 2). We also obtained the M/G1-related modules, [M/G1, RLM1] and [MCM1, RLM1]. RLM1 is a type II Mcm1-Arg80-Deficiens- Serum response factor (MADS)-box transcription factor that has been shown to be responsible for yeast cell wall organization and biogenesis (32) and may be involved in regulation of the M/G1 phase. Moreover, MCM1 is a type I MADS box protein, suggesting that RLM1 could interact with the N-terminal (NT) arm of MCM1. However, less evidence exists for the role of RLM1 as a cofactor recruited by the MCM1 NT arm (33)..

In sum, except for the case of the G2/M-specific modules, most of the phase-specific and phase-associate modules predicted by our integrative method are strongly consistent with published reports. When these phase-specific and phase-associate modules are connected to one another, the results may be drawn as the network shown in Fig. 2. This network is in good agreement with the result of other genomic analyses (15,27), and identifies and positions several additional modules, including [S, HIR1, HIR2], [MET4, CBF1] and [MET4, MET31], which are consistent with other experimental findings (23,24), but have not previously been placed within a network.

For Peer Review

Effects of the correlation threshold (P_c)

A less stringent threshold (P_c) will yield a larger number of modules and a larger number items per module. When we analyzed our datasets using a less stringent threshold (P_c <0.1) we obtained 59 (out of 1485) expected modules, which was more than we had obtained with the higher stringency threshold (P_c <0.01; yielding 43 modules). Notably, the lower-stringency threshold yielded two modules involved in the M/G1 phase, [M/G1, SWI5, ACE2] and [M/G1, STE12, DIG1], which were highly supported by the literature, but were not obtained with the higher stringency threshold (see Table 2).

STE12 and its inhibitor, DIG1, are involved in the regulation of mating-specific genes and pseudohyphal growth in the M/G1 phase (34), in a process that may or may not involve DIG1/DIG2 targeting by Fus3 (26). Thus, the [M/G1, STE12, DIG1] module identified under the less stringent threshold (P_c <0.1) is supported in the literature. Although DIG2 was not among the 113 TFs examined in the present work (15).

ACE2 and SWI5 are TFs that function at the M/G1 boundary (27,30). Under the stringent threshold (P_c <0.01), we obtained modules indicating that SWI5 and ACE2 are both involved in the M/G1 phase. However, under the less stringent threshold (P_c <0.1), these TFs formed module [M/G1, SWI5, ACE2], providing the additional information that SWI5 and ACE2 cooperate to regulate transcription in the M/G1 phase.

For Peer Review

Discussion

In the present study, we used a novel computational method to analyze combinatorial modules involved in the yeast cell cycle, by integrating data from previous ChIP and microarray expression studies and searching for correlations among TFs or among TFs and a phase of the cell cycle. Our results allowed us to propose a new regulation model describing the possible intersections of TFs involved in the G1, S, S/G2, G2/M, and M/G1 cell cycle phases (see Fig. 2).

Many previous methods for identifying associations and combinatorial interactions among cell-cycle regulators have been based solely on using gene expression data to either group co-regulated genes according to their expression similarities (8,9), or to connect over-represented TFs with binding motifs in the regulatory regions of target genes (10-14). However, the accuracy of these methods depends on the strategy used for grouping co-regulated genes, and it is often difficult to identify cooperativity among multiple TFs based on the observation of over-represented motifs.

Tsai *et al*. (3) used statistical methods (ANOVA) to study synergistic pairs of TFs in the yeast cell cycle by integrating ChIP-chip data from Harbison *et al*. (35) and microarray data from Spellman *et al*. (16). Here, we used published ChIP microarray data for most of the yeast TFs from Lee *et al.* (15) instead and used the same expression microarray data to define a novel measure of correlation among items (TF-targeted genes and cell-cycle phases) to extract TF regulatory modules with more biological relevance. We obtained 43 out of 909 assembled modules with stringent threshold of $Pc<0.01$ and 59 out of 1485 assembled modules with less stringency of $Pc<0.1$. Comparison reveals that our results are consistent with those of Tsai *et al*. (3), even though our study was based on a different ChIP dataset. However, our method using an

For Peer Review

Apriori algorithm yielded more significant combinations in the form of phase-associate modules which could provide more information than previous studies. For example, our method identified phase-associate modules such as [MBP1, STB1, SWI4], [MET4, CBF1] and [MET4, MET31], all of which contain elements that have been indirectly related to the yeast cell cycle phases in published reports (19, 25). In addition, our results also strongly agree with those of Kato *et al*. (1), who used 2 × 2 contingency table test with Yates correction to analyze a ChIP dataset in combination with a yeast microarray expression data in an effort to identify combinatorial regulation of TFs and corresponding binding motifs.

Previous studies have shown that some TF-targeted genes are co-regulated only under certain experimental conditions in cell cycle microarray data, but behave almost independently under other experimental conditions. Thus, the classic approach in using global expression patterns to identify similarities among module intersections is limited because it uses a limited amount of information. Taking this into account, we used phase-specific genes previously classified according to transcriptional oscillation through one peak per cell cycle (16), and applied these genes for the identification of associations between known TFs and specific phases of the yeast cell cycle. Our method of integrating genome-wide gene expression data with ChIP data successfully identified regulatory modules with biological relevance in the yeast cell cycle, and may be easily applied to other biological processes. For example, in addition to integrating ChIP and expression microarray data, our method has sufficient power to integrate other genomic resources, such as genome-wide protein-protein interaction data and gene ontology databases. Furthermore, we believe that integrated approaches will eliminate noises contained in each individual dataset, increasing the accuracy of our results compared to those obtained through non-integrated approaches.

For Peer Review

Our analysis of associations among the ChIP dataset also allowed us to identify a number of modules that were unrelated to the cell cycle phase, including [GAT3, PDR1, SWI5, YAP5], [HAP5, HAP3, HAP2], [HAP4, HAP3, HAP2] and [YAP1, SKN7] (see Supplementary Materials). GAT3, PDR1, SWI5 and YAP5 have previously been associated with metabolism (4), HAP2, HAP3, HAP4 and HAP5 are four subunits of the *Saccharomyces cerevisiae* CCAAT-binding factor (36), and YAP1 and SKN7 are thought to be involved in the oxidative stress response (37). Thus, our findings are strongly consistent with published evidence. However, it would be useful in future studies to integrate our data with a gene ontology dataset, which could allow us to directly assign these modules to the appropriate biological processes.

Several previous studies have sought to identify cooperativity among TFs (1-7). However, the studies by Banerjee and Zhang (2), Tsai *et al*. (3), Chang *et al*. (6) and Ruan and Zhang (7) identified only synergistic pairs of TFs, not modules containing multiple TFs, which are necessary to understand true cooperativity among TFs. In addition, the studies by Banerjee *et al*. (2), Nagamine *et al*. (4) and Chang *et al*. (6) did not assign their identified cooperative TFs to relevant biological processes. Thus, the previous works were limited in terms of their generated information. For example, experimental evidence has shown that MCM1 and FKH1 cooperate to recruit the NDD1 protein in the G2 phase (29), and our method allowed us to identify the largest correlative module [MCM1, NDD1, FKH2] and assign it to the G2 phase, thus providing far more information than obtained in the previous studies.

It should be noted that our predicted modules were all dependent on P_c. With a less stringent P_c, [more modules were identified and] the number of elements within each module is greater, making it possible to identify modules that might not be obtained under the more stringent threshold. For example, the module, [M/G1, SWI5, ACE2] is

For Peer Review

consistent with experimental evidence and is more informative than modules [M/G1, SWI5] and [M/G1, ACE2], but the larger module was obtained only under the less stringent threshold in our present study (P_c<0.1 versus P_c<0.01).

Conclusion

We herein describe an Apriori-like approach for discovering cooperative combinations of multiple TFs even when their binding motifs are not distinct. Our method has the following advantages: (i) it can integrate multiple genome-wide datasets, thus eliminating noise; (ii) it can detect modules containing multiple items instead of only synergistic pairs; and (iii) it can directly assign the identified modules to the appropriate cell cycle phase if a phase-related item is contained within the module. We used this method to integrate genome-wide gene expression data and chromatin immunoprecipitation (ChIP) data, and achieved high modeling accuracy pertaining to the yeast cell cycle. We identified numerous phase-specific and phase-associate modules that were strongly consistent with previous experimental findings, and additionally identified previously unknown modules that may be worthy of further study. With new genomic resources becoming available each year, we believe that our new, generalizable integrative method will prove highly useful for analysis of large-scale transcriptional regulatory networks.

For Peer Review

References

1. Kato, M., Hata, N., Banerjee, N., Futcher, B. and Zhang, M.Q. (2004) Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol*, **5**, R56.
2. Banerjee, N. and Zhang, M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res*, **31**, 7024-7031.
3. Tsai, H.K., Lu, H.H. and Li, W.H. (2005) Statistical methods for identifying yeast cell cycle transcription factors. *Proc Natl Acad Sci U S A*, **102**, 13532-13537.
4. Nagamine, N., Kawada, Y. and Sakakibara, Y. (2005) Identifying cooperative transcriptional regulations using protein-protein interactions. *Nucleic Acids Res*, **33**, 4828-4837.
5. Lemmens, K., Dhollander, T., De Bie, T., Monsieurs, P., Engelen, K., Smets, B., Winderickx, J., De Moor, B. and Marchal, K. (2006) Inferring transcriptional modules from ChIP-chip, motif and microarray data. *Genome Biol*, **7**, R37.
6. Chang, Y.H., Wang, Y.C. and Chen, B.S. (2006) Identification of transcription factor cooperativity via stochastic system model. *Bioinformatics*, **22**, 2276-2282.
7. Ruan, J. and Zhang, W. (2004), *Technical Report 43*. Department of Computer Science and Engineering,Washington University in St Louis.
8. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, **95**, 14863-14868.
9. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, **96**, 2907-2912.
10. Bannai, H., Inenaga, S., Shinohara, A., Takeda, M. and Miyano, S. (2004) Efficiently finding regulatory elements using correlation with gene expression. *J Bioinform Comput Biol*, **2**, 273-288.
11. Pizzi, C., Bortoluzzi, S., Bisognin, A., Coppe, A. and Danieli, G.A. (2005) Detecting seeded motifs in DNA sequences. *Nucleic Acids Res*, **33**, e135.
12. Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, **2**, 28-36.
13. Roth, F.P., Hughes, J.D., Estep, P.W. and Church, G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol*, **16**, 939-945.
14. Liu, X.S., Brutlag, D.L. and Liu, J.S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol*, **20**, 835-839.
15. Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I. *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799-804.
16. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by

For Peer Review

microarray hybridization. *Mol Biol Cell*, **9**, 3273-3297.
17. Agrawal, R. and Srikant, R. (1994), *Readings in database systems (3rd ed.)*. Morgan Kaufmann Publishers Inc., pp. 580-592.
18. Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M. and Brown, P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533-538.
19. Koch, C., Moll, T., Neuberg, M., Ahorn, H. and Nasmyth, K. (1993) A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. *Science*, **261**, 1551-1557.
20. Costanzo, M., Schub, O. and Andrews, B. (2003) G1 transcription factors are differentially regulated in Saccharomyces cerevisiae by the Swi6-binding protein Stb1. *Mol Cell Biol*, **23**, 5064-5077.
21. Ho, Y., Costanzo, M., Moore, L., Kobayashi, R. and Andrews, B.J. (1999) Regulation of transcription at the Saccharomyces cerevisiae start transition by Stb1, a Swi6-binding protein. *Mol Cell Biol*, **19**, 5267-5278.
22. Horak, C.E., Luscombe, N.M., Qian, J., Bertone, P., Piccirrillo, S., Gerstein, M. and Snyder, M. (2002) Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. *Genes Dev*, **16**, 3017-3033.
23. Spector, M.S., Raff, A., DeSilva, H., Lee, K. and Osley, M.A. (1997) Hir1p and Hir2p function as transcriptional corepressors to regulate histone gene transcription in the Saccharomyces cerevisiae cell cycle. *Mol Cell Biol*, **17**, 545-552.
24. Blaiseau, P.L. and Thomas, D. (1998) Multiple transcriptional activation complexes tether the yeast activator Met4 to DNA. *Embo J*, **17**, 6327-6336.
25. Thomas, D. and Surdin-Kerjan, Y. (1997) Metabolism of sulfur amino acids in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev*, **61**, 503-532.
26. Wittenberg, C. and Reed, S.I. (2005) Cell cycle-dependent transcription in yeast: promoters, transcription factors, and transcriptomes. *Oncogene*, **24**, 2746-2755.
27. Simon, I., Barnett, J., Hannett, N., Harbison, C.T., Rinaldi, N.J., Volkert, T.L., Wyrick, J.J., Zeitlinger, J., Gifford, D.K., Jaakkola, T.S. *et al.* (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697-708.
28. Breeden, L.L. (2003) Periodic transcription: a cycle within a cycle. *Curr Biol*, **13**, R31-38.
29. Koranda, M., Schleiffer, A., Endler, L. and Ammerer, G. (2000) Forkhead-like transcription factors recruit Ndd1 to the chromatin of G2/M-specific promoters. *Nature*, **406**, 94-98.
30. Doolin, M.T., Johnson, A.L., Johnston, L.H. and Butler, G. (2001) Overlapping and distinct roles of the duplicated yeast transcription factors Ace2p and Swi5p. *Mol Microbiol*, **40**, 422-432.
31. Pramila, T., Miles, S., GuhaThakurta, D., Jemiolo, D. and Breeden, L.L. (2002) Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle. *Genes Dev*, **16**, 3034-3045.
32. Jung, U.S. and Levin, D.E. (1999) Genome-wide analysis of gene expression regulated by the yeast cell wall integrity signalling pathway. *Mol Microbiol*, **34**, 1049-1057.
33. Abraham, D.S. and Vershon, A.K. (2005) N-terminal arm of Mcm1 is required for transcription of a subset of genes involved in maintenance of the cell wall. *Eukaryot Cell*, **4**, 1808-1819.

For Peer Review

34. Olson, K.A., Nelson, C., Tai, G., Hung, W., Yong, C., Astell, C. and Sadowski, I. (2000) Two regulators of Ste12p inhibit pheromone-responsive transcription by separate mechanisms. *Mol Cell Biol*, **20**, 4199-4209.
35. Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99-104.
36. Tanaka, A., Kato, M., Nagase, T., Kobayashi, T. and Tsukagoshi, N. (2002) Isolation of genes encoding novel transcription factors which interact with the Hap complex from Aspergillus species. *Biochim Biophys Acta*, **1576**, 176-182.
37. Lee, J., Godon, C., Lagniel, G., Spector, D., Garin, J., Labarre, J. and Toledano, M.B. (1999) Yap1 and Skn7 control two specialized oxidative stress response regulons in yeast. *J Biol Chem*, **274**, 16040-16046.
38. Futcher, B. (2002) Transcriptional regulatory networks and the yeast cell cycle. *Curr Opin Cell Biol*, **14**, 676-683.

For Peer Review

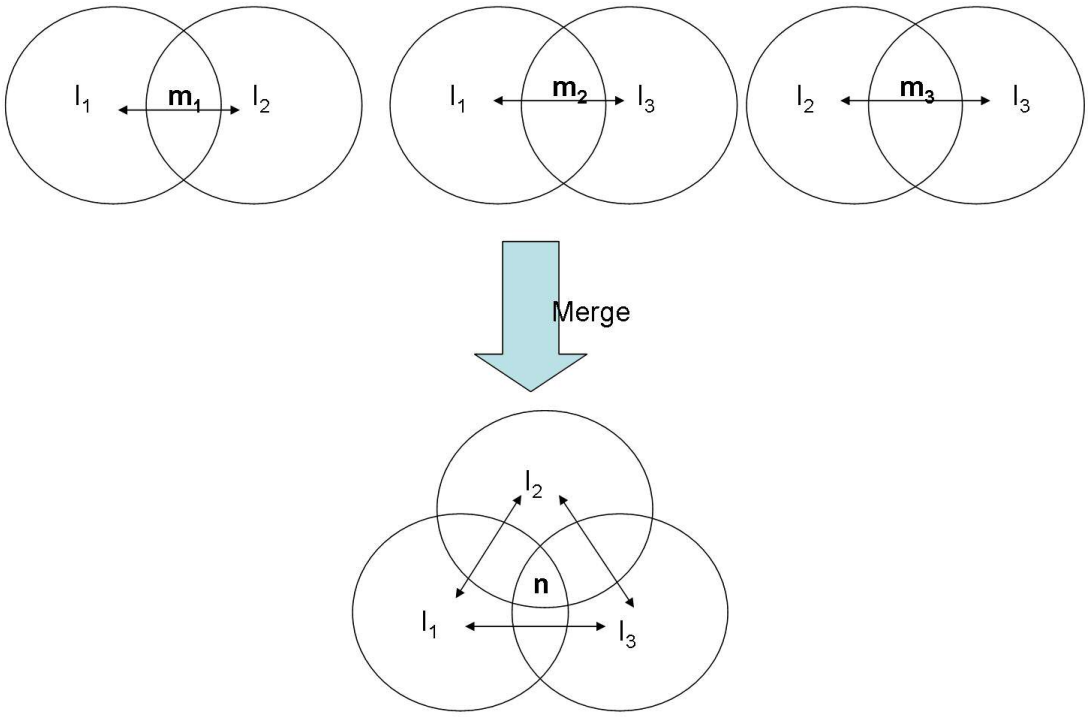


Fig. 1. A schematic depiction for the formation of 3-order extended modules from merging 2-order modules. We know that [I_1, I_2], [I_1, I_3] and [I_2, I_3] are correlative 2-order modules, as confirmed by using m_1, m_2 and m_3 intersection genes as m to calculate equation (1). We may then merge [I_1, I_2], [I_1, I_3] and [I_2, I_3] into the larger 3-order module, [I_1, I_2, I_3], but we then must assess the intersecting genes in the new module, to see whether they still correlate with one another. We use n intersection genes as m for equation (1) to calculate pairwise correlation between items among this module. If all pairwise correlations in this module are less than P_c, we can regard this as a correlative module.

For Peer Review

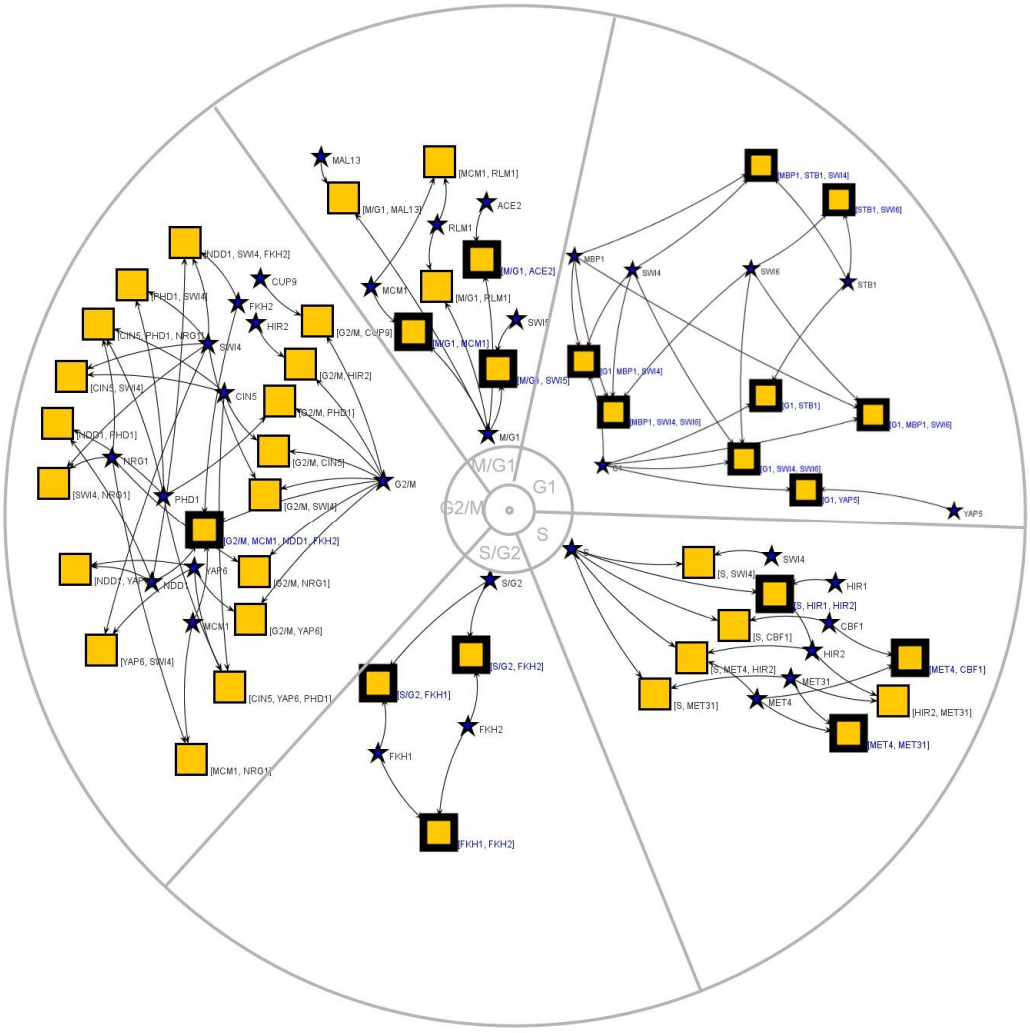


Fig. 2. An integrated yeast cell-cycle map organized by interactions of overall predicted items. Each item (a TF or cell cycle phase) is shown as a star, and modules are given as squares. A line connecting an item to a module indicates that our analysis suggests the item should be included in the module. Each item may be included in more than one module. Modules (squares) emphasized with black frames are supported by published evidence. Most of the modules are consistent with previous findings, though lack of previous support for most of the G2/M-related modules.

For Review

Table 1. The predicted phase-specific and phase-associate modules ($P_b^a < 0.01$, $P_c^b < 0.01$, $I_{min}^c = 5$).

Regulator of modules	Known protein interactions	Literature evidences
G1 Association:		
[G1, STB1]	[G1, STB1]	(21)
[G1, YAP5]	[G1, YAP5]	(22)
[G1, MBP1, SWI4]	[G1, MBP1, SWI6]	(18,19)
[G1, MBP1, SWI6]	[G1, MBP1],[G1,SWI4]	(18,19)
[G1, SWI4, SWI6]	[G1, SWI4, SWI6]	(18,19)
[STB1, SWI6]	[STB1, SWI6]	(21)
[MBP1, STB1, SWI4]	[MBP1, STB1, SWI4]	(21)
[MBP1, SWI4, SWI6]	[MBP1, SWI6],[SWI4, SWI6]	(18)
S Association:		
[S, CBF1]		
[S, SWI4]		
[S, MET31]		
[S, MET4, HIR2]	[S, HIR2]	(23)
[S, HIR1, HIR2]	[S, HIR1, HIR2]	(23)
[HIR2, MET31]		
[MET4, CBF1]	[MET4, CBF1]	(24)
[MET4, MET31]	[MET4, MET31]	(24)
S/G2 Association:		
[S/G2, FKH1]	[S/G2, FKH1]	(15,27,38)
[S/G2, FKH2]	[S/G2, FKH2]	(15,27,38)
[FKH1, FKH2]	[FKH1, FKH2]	(28)
G2/M Association:		
[G2/M, CIN5]		
[G2/M, YAP6]		
[G2/M, HIR2]		
[G2/M, PHD1]		
[G2/M, SWI4]		
[G2/M, CUP9]		
[G2/M, NRG1]		
[G2/M, MCM1, NDD1, FKH2]	[G2/M, MCM1, NDD1, FKH2]	(29)
[MCM1, NRG1]		
[CIN5, SWI4]		
[NDD1, YAP6]		
[NDD1, PHD1]		
[YAP6, SWI4]		
[PHD1, SWI4]		
[SWI4, NRG1]		
[CIN5, YAP6, PHD1]		
[CIN5, PHD1, NRG1]		
[NDD1, SWI4, FKH2]		
M/G1 Association:		
[M/G1, MCM1]	[M/G1, MCM1]	(27,31)
[M/G1, SWI5]	[M/G1, SWI5]	(27,30)
[M/G1, RLM1]		
[M/G1, MAL13]		
[M/G1, ACE2]	[M/G1, ACE2]	(27,30)
[MCM1, RLM1]		

For Peer Review

Most of the phase-specific modules are consistent with evidence found in the literature. However, only one G2/M-specific module, [G2/M, MCM1, NDD1, FKH2], is strongly support by previous reports.

[a] P_b, *P*-value for TF binding to chromatin, as described by Lee *et al*. (15).

[b] P_c, threshold for the correlation between items. Pairwise correlations between items in a module, as calculated by equation (1), must be less than P_c for the module to be correlative.

[c] I_{min}, threshold for the number of interactions. The listed modules must have more than I_{min}. The minimal number of interactions was set to be 5 in the study.

For Peer Review

Table 2. Predicted phase-specific modules involved in the M/G1 phase (P_b[a] < 0.01, P_c[b] < 0.1, I_{min}[c] = 5).

Regulator of modules	Known protein interactions	Literature evidences
[M/G1, HAP4]		
[M/G1, MCM1]	[M/G1, MCM1]	(27,31)
[M/G1, GAT3]		
[M/G1, SUM1]		
[M/G1, PDR1]		
[M/G1, PHD1]		
[M/G1, SMP1]		
[M/G1, RLM1]		
[M/G1, MAL13]		
[M/G1, STE12, DIG1]	[M/G1, STE12, DIG1]	(26,34)
[M/G1, MSN4, RGM1]		
[M/G1, SWI5, ACE2]	[M/G1, SWI5, ACE2]	(27,30)

If we used a less stringent threshold (^bP_c < 0.1 instead of < 0.01), we obtained additional modules, including [M/G1, SWI5, ACE2] and [M/G1, STE12, DIG1].

[a] P_b, *P*-value for TF binding to chromatin, as described by Lee *et al*. (15).

[b] P_c, threshold for the correlation between items. Pairwise correlations between items in a module, as calculated by equation (1), must be less than P_c for the module to be correlative.

[c] I_{min}, threshold for the number of interactions. The listed modules must have more than I_{min}.

For Peer Review

Algorithm 1. Algorithm for finding the largest correlative modules

1: Scan the transactions $t \in T$ to find correlative 1-order modules in C_1;

2: for ($k = 2$; $C_{k-1} \neq \varnothing$; $k++$) do begin

3: Generate extended k-order modules in E_k by merging correlative $(k-1)$-order modules in C_{k-1};

4: if ($E_k = \varnothing$) do begin

5: Extract all $k-1$-order correlative modules in C_{k-1} as largest correlative modules;

6: break;

7: end

8: Prune extended k-order modules whose sub-modules are not correlative;

9: Scan the transactions $t \in T$ to count the co-occurrences of modules as m for equation (1) in extended k-order modules which survived the pruning process;

10: Save the correlative k-order modules into C_k, a subset of the extended k-order modules with pairwise correlations calculated by equation (1) no bigger than P_c;

11: Extract $k-1$-order largest correlative modules from correlative $(k-1)$-order modules in C_{k-1} if it is hard to form any correlative k-order modules in C_k with its associated $(k-1)$-order modules;

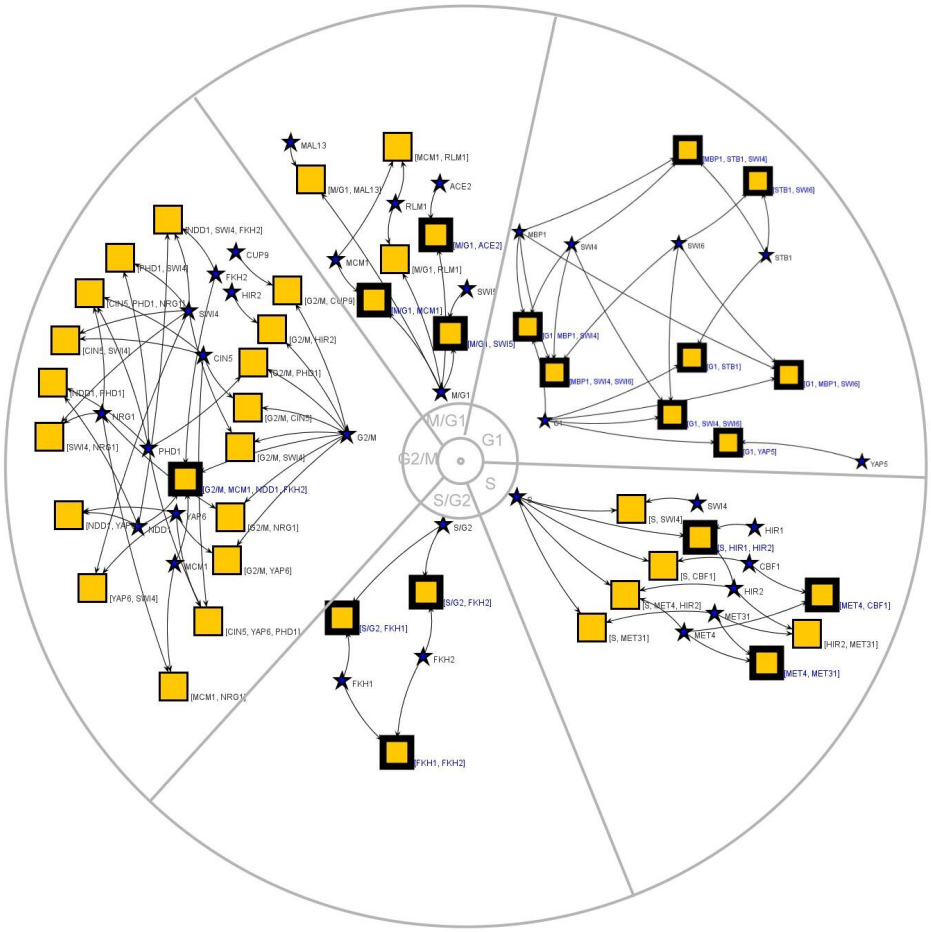
12: end

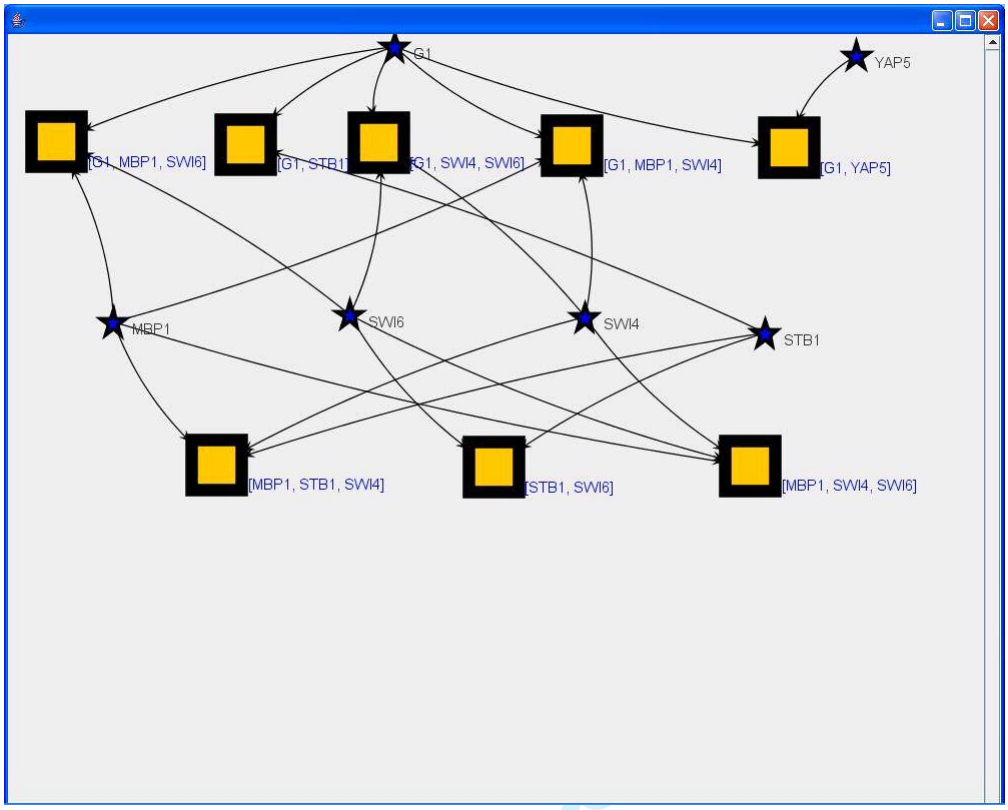
For Peer Review

Acknowledgments

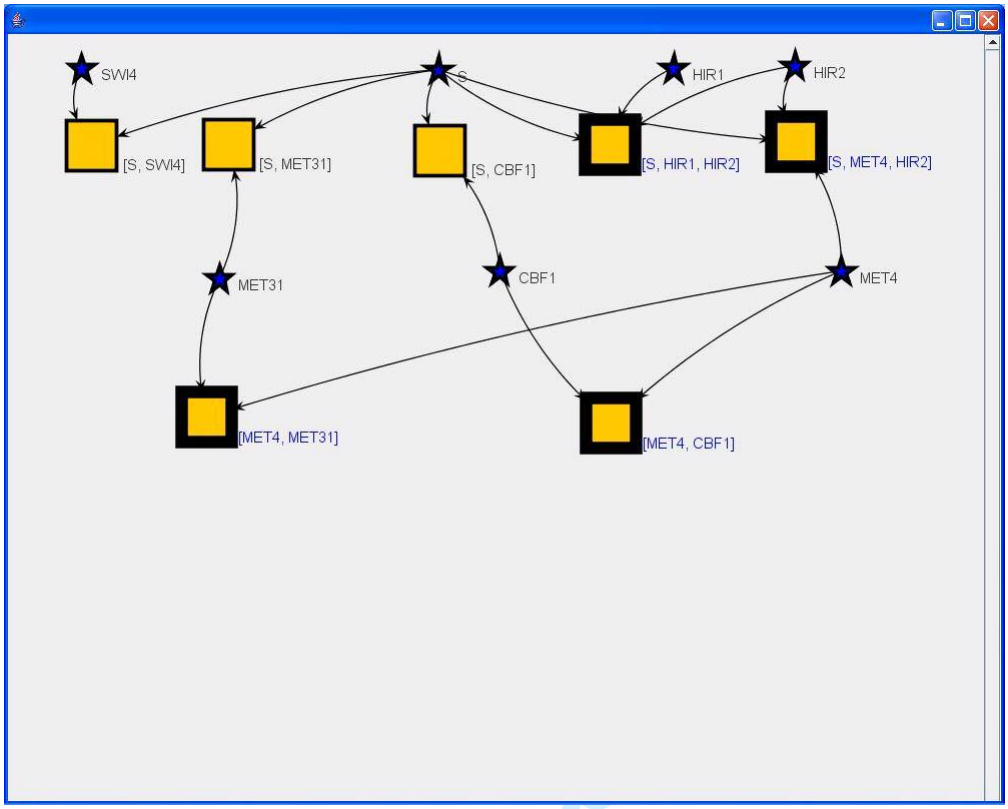
The work was supported by a Chang Gung Research Grant (CMRPD140041) and by the Molecular Medicine Research Center, Chang Gung University, Taiwan.

For Peer Review

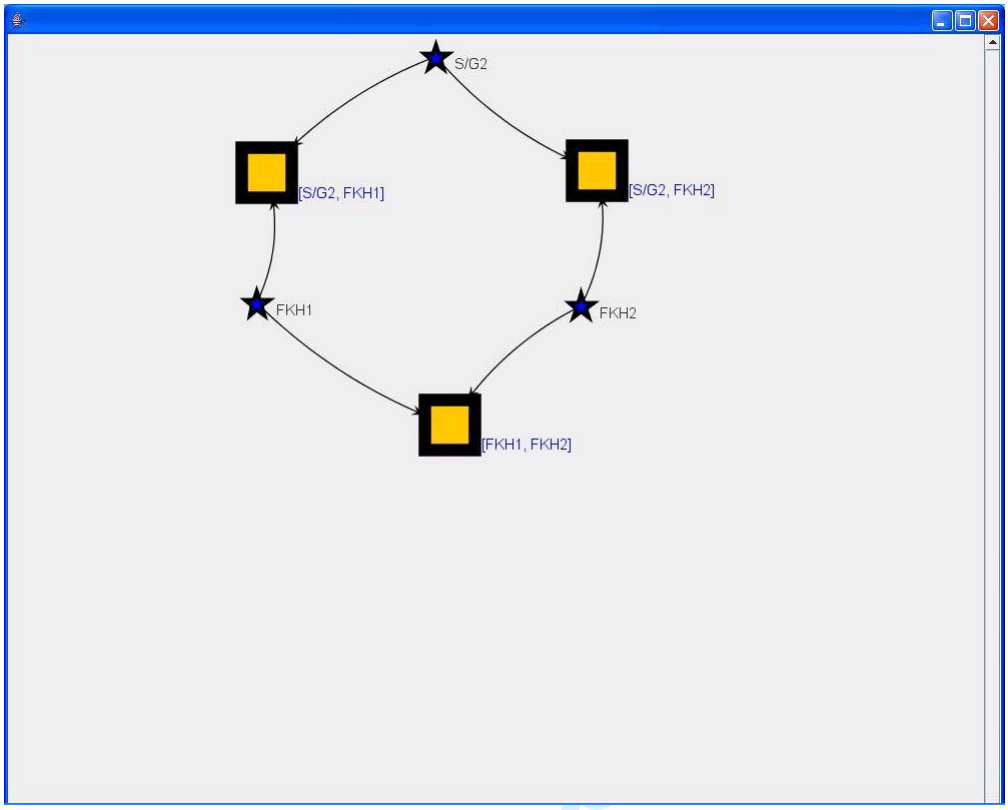

M/G1
G1
G2/M
S
S/G2
MAL13
[M/G1, MAL13]
[MCM1, RLM1]
ACE2
RLM1
[M/G1, ACE2]
[M/G1, RLM1]
MCM1
[M/G1, MCM1]
SWI5
[M/G1, SWI5]
M/G1
[NDD1, SWI4, FKH2]
CUP9
PHD1, SWI4]
FKH2
HIR2
[CIN5, PHD1, NRG1]
SWI4
[G2/M, CUP9]
[G2/M, HIR2]
[CIN5, SWI4]
CIN5
[G2/M, PHD1]
[NDD1, PHD1]
NRG1
[G2/M, CIN5]
[SWI4, NRG1]
PHD1
G2/M
[G2/M, SWI4]
[G2/M, MCM1, NDD1, FKH2]
YAP6
[G2/M, NRG1]
[NDD1, YAP6, MDD1]
MCM1
[G2/M, YAP6]
[YAP6, SWI4]
[CIN5, YAP6, PHD1]
[MCM1, NRG1]
S/G2
[S/G2, FKH2]
[S/G2, FKH1]
FKH2
FKH1
[FKH1, FKH2]
MBP1
SWI4
SWI6
STB1
[MBP1, STB1, SWI4]
[STB1, SWI6]
[G1, MBP1, SWI4]
[MBP1, SWI4, SWI6]
[G1, STB1]
[G1, MBP1, SWI6]
[G1, SWI4, SWI6]
[G1, YAP5]
YAP5
[S, SWI4]
SWI4
HIR1
[S, HIR1, HIR2]
[S, CBF1]
CBF1
HIR2
[S, MET4, HIR2]
MET31
MET4
[S, MET31]
[MET4, CBF1]
[HIR2, MET31]
[MET4, MET31]


G1
YAP5
[G1, MBP1, SWI6]
[G1, STB1]
[G1, SWI4, SWI6]
[G1, MBP1, SWI4]
[G1, YAP5]
MBP1
SWI6
SWI4
STB1
[MBP1, STB1, SWI4]
[STB1, SWI6]
[MBP1, SWI4, SWI6]

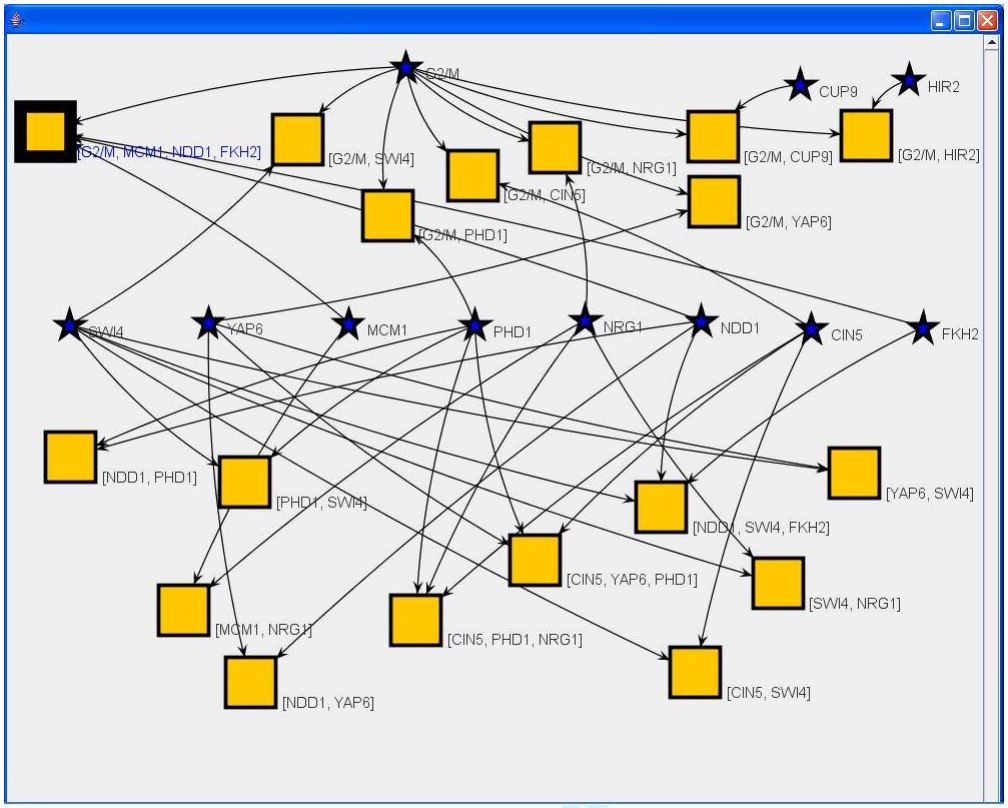
For Review


SWI4
S
HIR1
HIR2
[S, SWI4]
[S, MET31]
[S, CBF1]
[S, HIR1, HIR2]
[S, MET4, HIR2]
MET31
CBF1
MET4
[MET4, MET31]
[MET4, CBF1]

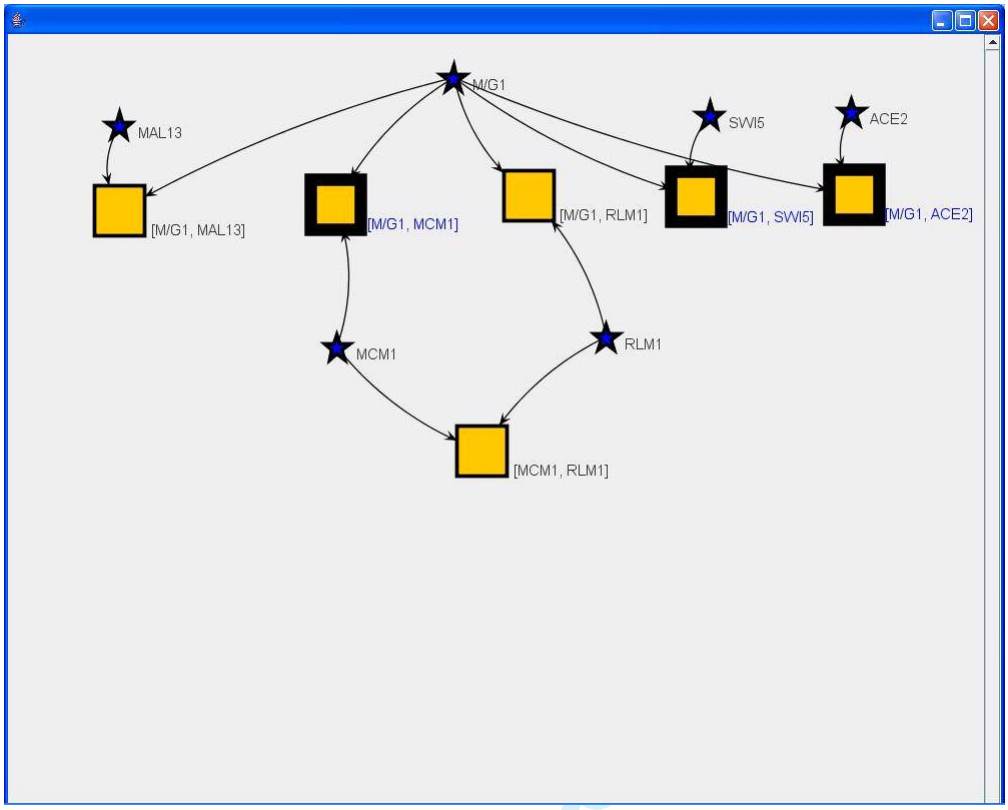
For Review



For Review


G2/M
CUP9
HIR2
[G2/M, MCM1, NDD1, FKH2]
[G2/M, SWI4]
[G2/M, NRG1]
[G2/M, CUP9]
[G2/M, HIR2]
[G2/M, CIN5]
[G2/M, YAP6]
[G2/M, PHD1]
SWI4
YAP6
MCM1
PHD1
NRG1
NDD1
CIN5
FKH2
[NDD1, PHD1]
[PHD1, SWI4]
[YAP6, SWI4]
[NDD1, SWI4, FKH2]
[CIN5, YAP6, PHD1]
[SWI4, NRG1]
[MCM1, NRG1]
[CIN5, PHD1, NRG1]
[NDD1, YAP6]
[CIN5, SWI4]

For Review


M/G1
MAL13
SWI5
ACE2
[M/G1, MAL13]
[M/G1, MCM1]
[M/G1, RLM1]
[M/G1, SWI5]
[M/G1, ACE2]
MCM1
RLM1
[MCM1, RLM1]

For Review