OXFORD JOURNALS
Nucleic Acids Research

Maximization of negative correlations in time-course gene expression data for enhancing understanding of molecular pathways

Journal:	*Nucleic Acids Research*
Manuscript ID:	NAR-01155-Met-S-2009.R2
Manuscript Type:	3 Methods Manuscript (Online Publication) - US Editorial Office
Key Words:	maximization of negative correlations, time-course gene expression, molecular pathway, bi-clique

Published online XX XXXX 2009 *Nucleic Acids Research, 2009, Vol. XX, No. XX* **1–12**
doi:10.1093/nar/gkn000

Maximization of negative correlations in time-course gene expression data for enhancing understanding of molecular pathways

Tao Zeng, Jinyan Li *

School of Computer Engineering & Bioinformatics Research Center, Nanyang Technological University, Nanyang Avenue, Singapore 639798

Received XXXX, 2009; Revised XXXX, 2009; Accepted XXXX, 2009

ABSTRACT

Positive correlation can be diversely instantiated as shifting, scaling, or geometric pattern, and it has been extensively explored for time-course gene expression data and pathway analysis. Recently, biological studies emerge a trend focusing on the notion of negative correlations such as opposite expression patterns, complementary patterns, and self negative regulation of transcription factors. These biological ideas and primitive observations motivate us to formulate and investigate the problem of maximizing negative correlations. The objective is to discover all maximal negative correlations of statistical and biological significance from time-course gene expression data for enhancing our understanding of molecular pathways.

Given a gene expression matrix, a maximal negative correlation is defined as an activation-inhibition two-way expression pattern (AIE pattern). We propose a parameter-free algorithm to enumerate the complete set of AIE patterns from a data set. This algorithm can identify significant negative correlations that cannot be identified by the traditional clustering/biclustering methods. To demonstrate the biological usefulness of AIE patterns in the analysis of molecular pathways, we conducted deep case studies for AIE patterns identified from Yeast cell cycle data sets. In particular, in the analysis of the Lysine biosynthesis pathway, new regulation modules and pathway components were inferred according to a significant negative correlation which is likely caused by a co-regulation of the transcription factors at the higher layer of the biological network. We conjecture that maximal negative correlations between genes are actually a common characteristic in molecular pathways, which can provide insights into the cell stress response study, drug response evaluation, etc.

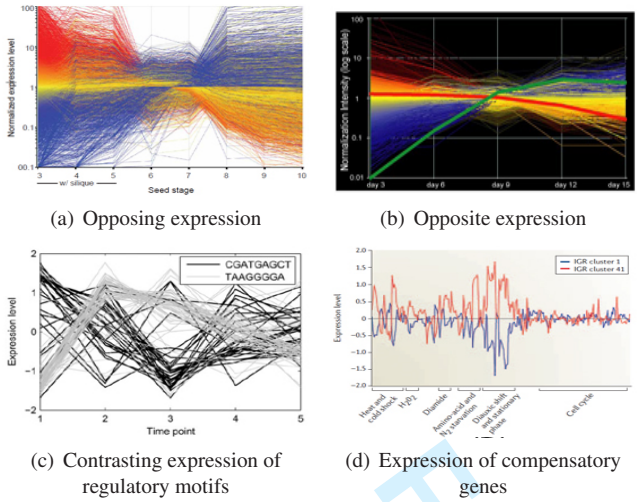
INTRODUCTION

A molecular pathway is referred to as a series of actions among molecules in a cell leading to a certain end point of cell function. Pathway identification is usually aimed to uncover all biological molecules participating in the same functional pipeline, which may include DNA/gene, miRNA, protein or metal ion, etc. As DNA and protein play the major roles in a pathway, gene and protein's indirect relations are of paramount importance for detecting and analyzing molecular pathways.

Gene expression data, especially time-course gene expression data, have been widely used to explore various relationships of the genes in the pathways, with the particular focus on the *positive correlations*. For example, Segal *et al.* proposed to identify new pathways by assuming that most genes in the same pathway can exhibit a *similar* gene expression profile, and their proteins often interact (1); Multiplicative patterns and scaling patterns have been also used to describe the expression profiles of the genes in the same pathway (2, 3, 4); Co-regulation patterns, additive expression patterns or shifting patterns, have been conceptualized to detect regulatory modules from gene expression data (5, 6); Further, geometric patterns based on trigonometric functions are believed to be related to circular regulation processes (7). Here, concepts such as profile similarity, shifting pattern, scaling pattern, or geometric pattern are all concentrated on positive correlations among genes, implying that genes with expression homogeneity are possible to have the same biological function.

In this work, we are interested in *negative correlation*. It is also an important relationship among genes, and it has been previously observed in many biological processes. Schmid *et al.* had a study on development expression patterns for large gene families of *Arabidopsis thaliana* (8); they highlighted two groups of genes showing an *opposing* expression trends from an early seed development stage to a late stage (Fig. 1(a)). In a study of expression patterns in the chondrogenic differentiation, James *et al.* had a careful analysis on a 15-day temporal gene expression data of a *Mouse* micromass culture system (9), and they reported an interesting example on two groups of genes displaying an *opposite* expression pattern. In that example, transcription factors Sox9 showed high expression levels before day 6,

*To whom correspondence should be addressed. Tel: +65 67906253; Fax: +65 6792 6559; Email: jyli@ntu.edu.sg



(a) Opposing expression (b) Opposite expression

(c) Contrasting expression of regulatory motifs (d) Expression of compensatory genes

Figure 1. Notion of negative correlations in biological studies.

then their expression decreased by about half; while, Ibsp transcripts showed low expression levels until day 12 and then had 1200-fold up-regulation (Fig. 1(b)). Chuang *et al.* introduced the notion of *complementary* gene expression patterns for inferring time-lagged genetic interactions (10), which is intended to capture contrasting expression patterns like the one that: when one gene's expression increases, the other gene's decreases, or vice versa. As another example, Stekel *et al.* proposed a method for modeling the so called *self negative regulation* of transcription factors (11), in which the product of one gene is assumed to regulate this gene's transcription factor in a feedback way.

Based on above biological ideas and primitive observations, we propose to formulate and systematically investigate the problem of *maximizing negative correlations*. The objective is to discover all maximal negative correlations from time-course gene expression data. A maximal negative correlation is defined as a pair of correlated *gene sets* where genes between the two sets must be negatively correlated in their expression over a time segment, and the number of genes and the number of time points in the segment are required to be maximized. Therefore, maximization of negative correlations, like detecting positive correlations, can expand and deepen down the analysis on molecular pathways. This is in a good agreement with a biological fact that: when two genes have a negatively correlated interaction, then the two group genes locating at the downstream of their participating pathways will also form a negative relationship (12). To capture such a pipeline of negative correlations spanning multiple time points from time-course gene expression data, maximization of negative correlations is a novel and effective attempt.

Maximization of negative correlations is extendable and potentially applicable to a wide range of pathway studies where biological molecular relationships can be mapped to gene expression correlations. Segal *et al.* were interested in motif extraction from sequences of promoters, and observed that the expression patterns of regulatory motifs with nucleotide variation were bi-polarly different (13). See one example of such expression patterns in Fig. 1(c). Shieh *et al.* proposed to study transcriptional compensation interactions or synthetic lethal pairs with the idea that: following some gene's mutation, its *compensatory* gene will be up-regulated (or down-regulated) (14). Millar *et al.* were interested in the whole genome pattern of histone lysine acetylation and methylation in *Yeast* to confirm a hypothesis that different combinations of histone modification sites are likely associated with specific and contrasting transcription behaviors (Fig. 1(d)) (15). They also pointed out that these patterns can exist in other organisms such as *Schizosaccharomyces pombe* genome, portions of the *Drosophila melanogaster* even *Human* genomes. Recently, research results all show that such histone modification patterns are correlated with *Human* diseases (16, 17, 18). Therefore, through maximizing negative correlations, these biological applications can be certainly deepened down.

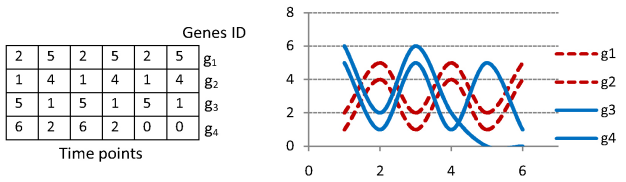
Given a gene expression data matrix, a maximal negative correlation can be viewed as an *activation-inhibition two-way expression pattern* (AIE pattern) where the two groups of genes exhibit such a behavior that when one group of genes is up-regulated, the other group is low-expressed, or vice versa, consistently at a continuous range of time points. Identifying a complete set of significant AIE patterns from gene expression data is computationally expensive. We design a new graph-based method for an exact and complete enumeration of AIE patterns with high efficiency. Our algorithm combines two mining strategies: a suffix-tree structure and a bi-clique approach for efficient search of the AIE patterns. To our best knowledge, there is no algorithm that can be specialized to identify AIE patterns. Clustering methods may be easy to find out activation-inhibition relation (7, 19), but there will be a lot of false positives, and local negative correlations under different specific time points cannot be identified. Biclustering can find gene expression patterns related to specific conditions, but it's difficult to mine large number of genes within negative relations (20, 21). The so called anti-correlated patterns (22, 23) are closely related to our AIE patterns, however, their mining algorithm cannot produce exactly AIE patterns (see a detailed comparison later). In our in-silicon evaluation, our method has been successfully applied to Yeast time-course gene expression data to reveal negative correlations in the molecular pathways of *Saccharomyces cerevisiae* for increasing the understanding of its biological mechanisms.

METHODS

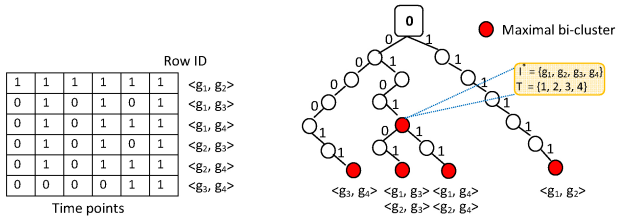
Let M be a time-course gene expression data set denoted as a triplet $M=(G,C,d)$, where $G=\{g_1,g_2,\cdots,g_n\}$ is a set of genes (rows), $C=\{c_1,c_2,\cdots,c_m\}$ is an ordered set of continuous time points (columns), and $d:G\times C\Rightarrow R$ is the level function by which $d(g_i,c_j)$ represents the expression level of gene g_i at time point c_j.

A *continuous subset* of $C=\{c_1,c_2,\cdots,c_m\}$ is an ordered subset of C with continuous time points. In other words, if $T=\{t_1,t_2,\cdots,t_k\}$ is a continuous subset of C, then $k\leq m$ and $t_i=c_{i+j}$, $i=1,...,k$, for some $j\in\{0,1,\cdots,m-k\}$.

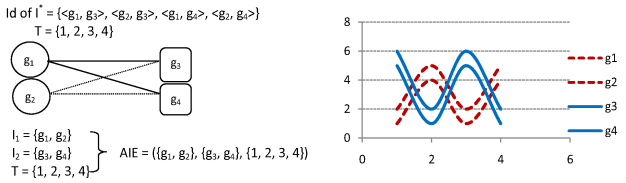
DEFINITION 1 (Activation-inhibition two-way expression pattern). *Let $X=(I,J,d)$ be a sub-matrix of a time-course gene expression data matrix $M=(G,C,d)$, where I is a subset of G and J is a continuous subset of C, X is defined as an*



(a) An original gene expression data matrix



(b) The dichotomy matrix and its suffix tree



(c) One bi-clique and its corresponding AIE pattern

Figure 2. A suffix-tree and a bi-clique search method are combined for AIE pattern mining

activation-inhibition two-way express pattern (AIE pattern) if I *can be divided into* I_1 *and* I_2 *such that at every time point* $j \in J$*, the expression levels of genes in* I_1 *and those in* I_2 *satisfy either*

$$d_{i_1,j} < d_{i_2,j}, \forall i_1 \in I_1, i_2 \in I_2$$

or

$$d_{i_1,j} > d_{i_2,j}, \forall i_1 \in I_1, i_2 \in I_2.$$

Suppose $X = (I,J,d)$ is an AIE pattern in a time-course gene expression data matrix. Usually, we require $|I|$ and $|J|$ are both maximal. This means that there are no genes or no time points that can be added into $X = (I,J,d)$ to maintain the conditions of the above definition.

Our method to enumerate a complete set of AIE patterns from a gene expression data set consists of three computational steps. The first step is to construct a *dichotomy matrix* based on the original data set, which captures and discretises the expression difference between every pair of genes at every time point. The second step is to transform the dichotomy matrix into another representation as suffix tree, and to extract the time-series biclusters from this tree in order to locate the time period when the genes possibly show a negative correlation. The third step is to construct a bi-partite graph according to the row ids (i.e., the gene pairs) from those biclusters, and to distinguish two groups of genes forming a bi-clique in such bi-partite graph.

DEFINITION 2 (Dichotomy Matrix). *Given a matrix* $M = (G,C,d)$*, its dichotomy matrix* $DM_{|G|\times(|G|-1)/2,|C|}(M)$ *is defined as:*

$$\begin{cases} DM_{i_1\times|I|-i_1\times(i_1+1)/2+(i_2-i_1)-1,j} = 0, \text{ if } d_{i_1,j} < d_{i_2,j} \\ DM_{i_1\times|I|-i_1\times(i_1+1)/2+(i_2-i_1)-1,j} = 1, \text{ if } d_{i_1,j} > d_{i_2,j} \end{cases}$$

where $i_1 < i_2$*. For each row* k *in* $DM(M)$*, if*

$$k = i_1\times|I| - i_1\times(i_1+1)/2 + (i_2-i_1) - 1, i_1 \in G, i_2 \in G$$

then, the row ID of this row is assigned as $< i_1, i_2 >$*.*

See an example of dichotomy matrix at the left panel of Fig. 2(b) which is derived from the gene expression matrix shown in the left panel of Fig. 2(a). In fact, each row of the dichotomy matrix can be considered as a 0-1 sequence, so a suffix tree of all 0-1 sequences/rows (the right panel of Fig. 2(b)) can be constructed in linear time (22). The depth of the nodes corresponds to the number of time points; the leaf nodes and the splitting nodes are marked with the row/gene pair IDs of the dichotomy matrix. Thus, every splitting node or leaf node is a maximal sub-matrix with every row identical in the dichotomy matrix. A proof about the relation between the nodes in a generalized suffix tree and the maximal biclusters with continuous columns can be found in (22).

For every splitting or leaf node, we construct a bi-partite graph using the gene pair IDs stored at the node. Suppose a node contains k number of gene pair IDs: $< i_1^1, i_2^1 >, \cdots, < i_1^k, i_2^k >$, then we denote $\{i_1^1, \cdots, i_1^k\}$ as the nodes at one side of the bi-partite graph, and denote $\{i_2^1, \cdots, i_2^k\}$ as the nodes at the other side of the bi-partite graph. Meanwhile, assign an edge between i_1^j and i_2^j for $j = 1,2,...,k$. See an example of bi-partite graph at the left panel of Fig. 2(c). Then, we enumerate all maximal bi-cliques from this bi-partite graph. Assume I_1 and I_2 are the two vertex sets of such a maximal bi-clique, then I_1 and I_2 are exactly the two non-overlapping gene groups for an AIE pattern whose time points are decided by the edge labels of the path leading to the splitting node or the leaf node from the root node in the suffix tree.

The pseudo-code of our algorithm is shown in **Algorithm** 1. The core sub-routine of this algorithm for the bi-clique mining is taken from (24). The whole algorithm can be divided into two parts and their computational complexities are analyzed as follows. The first part is a determinant routine to solve P-problem that discovers maximal row-identical sub matrices. The initiation of this algorithm (line 4) needs $O(G^2C)$ time and space; the suffix-tree construction (line 6) needs $O(G^2C)$ time and space (20); and the identification of maximal row-identical sub matrices from the suffix-tree (line 8) needs $O(G^2C^2)$ time. So the time and space cost of this determinant sub-routine are $O(G^2C^2)$ and $O(G^2C)$. While, the second part is an exhaustive pattern mining subroutine to solve an NP-problem that discovers all bi-cliques. The size of bi-partite graph input into bi-clique mining (line 9-17) is no more than $G \times C$, so under the worst condition, its time complexity is $O(G^2N)$ and space complexity is $O(G^2)$ (24). Here, N is the number of all maximal bi-cliques, or the number of all AIE patterns. It should be noted that, in the worst situation, the distribution of ***all*** potential AIE patterns will be dense in the

Algorithm 1 Suffix-tree and bi-clique two-stage method for detecting AIE patterns in time-course data

Require: a matrix M of size $G \times C$
Ensure: AIE patterns
1: **Int** $GPN = G(G-1)/2$
2: **Int** $GP[GPN][C]$
3: /*Produce dichotomy matrix of M at first stage*/
4: GP = **Dichotomy**(M)
5: /*Construct Suffix-tree at second stage*/
6: **SuffixTree** ST = **SuffixTree**(GP)
7: /*Extract row-identical sub-matrix Bic(I^*, J) in suffix-tree*/
8: **Bic** <ROW,COL> = **ExtractBic**(ST)
9: **for** each $<rows, cols>$ in <ROW, COL> **do**
10: /*Construct bi-partite graph B(I)*/
11: **Gragh** GG = **GenesGraph**($rows$)
12: /*Bi-clique mining at third stage*/
13: **Biclique** <ROW^G,COL^G> = **Biclique**(GG)
14: **for** each $<rows^G, cols^G>$ in <ROW^G, COL^G> **do**
15: **Output** $<rows^G, cols^G, cols>$ is an AIE pattern
16: **end for**
17: **end for**

original data. That means the N is so large that the time cost of AIE pattern mining will increase tremendously. Therefore, in actual applications, we will use the size, up-down index, and differential gap (see their definitions later) to limit the number of potential AIE patterns. The proposed algorithm will perform an exhaustive search for all AIE patterns under such specific constraints.

By definition, the negative correlation in an AIE pattern may go like the way that: at all time points in J, the expression of the genes in I_1 is always higher, or always lower, than those of genes in I_2. This is an extreme case of negative correlation. For the other cases, I_1's expression may be higher at the first time point, while turn to be lower at the second time point, and then come higher again than I_2's expression. This up-down trend at multiple time points can be measured by an ***up-down index*** value. This index can be used to categorize the expression trends exhibited by different AIE patterns. We also believe that this index value of an AIE pattern is related to the strength of a negative correlation.

DEFINITION 3 (Up-down index). *Let $X = (I, J, d)$ is an AIE pattern, its up-down vector U is a $|J| \times 1$ vector determined by:*

$$\forall j, U_j = \begin{cases} 0, \text{ if } d_{i_1,j} < d_{i_2,j}, \forall i_1 \in I_1, i_2 \in I_2 \\ 1, \text{ if } d_{i_1,j} > d_{i_2,j}, \forall i_1 \in I_1, i_2 \in I_2. \end{cases}$$

An up-down index value UI of X is defined as:

$$UI = \frac{|\{j \mid U_j \neq U_{j+1}, j = 1, ..., |J|-1\}|}{|J|-1}$$

Fig. 3 shows three examples of AIE pattern where the two genes marked with solid line are in one group, while the other two labeled as dotted line are in the second group. The AIE pattern in Fig. 3(a) has an up-down index value of $3/5$. While,

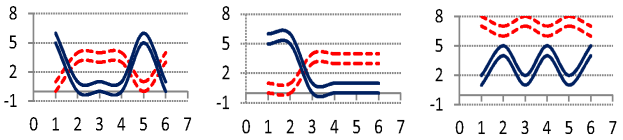


Figure 3. Examples of AIE patterns and their up-down trends.

the AIE pattern in Fig. 3(b) has an UI value as 1/5. In fact, the up-down index value of an AIE pattern is the up-down change frequency of one group's expression against the other group's over a continuous time points. For the case of negative correlation shown in Fig. 3(c), its up-down index value is 0. However, it still looks interesting. Therefore, we introduce a *generalized* Pearson's correlation coefficient to measure the negative correlation between two sets of variables. This is a finer correlation measurement compared to the up-down index.

Let I_1 and I_2 be two sets of gene variables with no overlapping. The generalized Pearson's correlation coefficient (noted as **R-value**) between I_1 and I_2 is denoted by $R(I_1, I_2)$, and it is calculated by

$$R(I_1, I_2) = \frac{\sum_{g_i \in I_1, g_j \in I_2} r(g_i, g_j)}{|I_1| * |I_2|}$$

where $r(g_i, g_j)$ is the Pearson's correlation coefficient between the two variables g_i and g_j. Note that the value of $R(I_1, I_2)$ is between -1 and 1, exactly with the same range as the conventional Pearson's correlation. The correlation is the most negative when the value -1 is reached, while on the other hand, it is the most positive when the value of 1 is approached. The generalized Pearson's correlation coefficient can also apply to a set I of gene variables to measure its inherent negative correlation if I can be properly divided into two sub-groups I_1 and I_2. In this work, such I_1 and I_2 are obtained by a hierarchical clustering approach.

Given an AIE pattern, the number of genes in one sub-group (I_1 or I_2) can be sometimes much smaller than the other group. So, we also introduce a ***group ratio*** index $GR(I_1, I_2)$ to measure the size balance between I_1 and I_2 for an AIE pattern, as calculated by

$$GR(I_1, I_2) = \frac{min(|I_1|, |I_2|)}{|I_1| + |I_2|}$$

To gain more insight into the negative correlation of an AIE pattern, we also examine how wide the expression between the two groups is. When the difference becomes wider, the expression behavior of the two groups are more distinct, thus the negative correlation may be more significant. To highlight this expression divergence between the two groups of genes I_1 and I_2, we define a ***differential gap*** to average the minimum difference of expressions between the two groups of genes, namely $\min_{i_1 \in I_1, i_2 \in I_2} |d_{i_1,j} - d_{i_2,j}|$ over multiple

Table 1. Different methods on negative correlation mining.

Methods	# of gene clusters from a data set (Alpha, Cdc15, Cdc28)	Average # of genes over all clusters in a data set (Alpha, Cdc15, Cdc28)	Average # of time points over all clusters in a data set (Alpha, Cdc15, Cdc28)
HCL-30	(26,14,30)	(21, 35, 17)	(18, 24, 17)
HCL-50	(47,34,50)	(11, 10, 10)	(18, 24, 17)
HCL-100	(99,82,100)	(5, 5, 5)	(18, 24, 17)
Kmeans-30	(30,30,30)	(17, 17, 17)	(18, 24, 17)
Kmeans-50	(50,48,50)	(10, 10, 10)	(18, 24, 17)
Kmeans-100	(99,89,94)	(5, 6, 5)	(18, 24, 17)
CC	(100,96,96)	(19, 10, 15)	(8, 6, 7)
OPSM	(3,2,4)	(27, 58, 23)	(10, 15, 9)
e-CCC-Biclustering	(43,70,28)	(14, 16, 12)	(10, 12, 10)
AIE	(105,165,151)	(11, 11, 11)	(10, 10, 10)

Here, **Average # of time points** indicates the average number of time points in the clusters or biclusters. Note that these values for HCL or Kmeans clusters are always the same.

time points. This differential gap is calculated by

$$G(I_1,I_2)=\frac{\sum_{j\in J}\min_{i_1\in I_1,i_2\in I_2}|d_{i_1,j}-d_{i_2,j}|}{|J|}$$

RESULTS

The data used in our evaluation is a time-course gene expression data repository related to Yeast. The raw data were published by the Yeast cell cycle analysis project (25). This project had acquired the expression measurements of 6179 genes involved in the Yeast cell cycle under three different conditions: *Alpha factor*-based synchronization, Cdc15-based synchronization and Cdc28-based synchronization. These measurements are actually the relative expressions against the control/background data; the zero expression represents the control level. We denote these three data sets simply as *Alpha factor*, *Cdc15* and *Cdc28* data set in this paper. The number of time points was set by the project as 18, 24, or 17 respectively for the *Alpha factor*, *Cdc15*, or *Cdc28* condition. The time point number 18 for the *Alpha factor* condition (similarly, 24 and 17 for the other two conditions) means that the RNA samples for the 6179 genes were collected at 18 time points starting at 0 to 7, 14, 21, ..., till the 119th minute, which covers two cell cycles. In our data preprocessing, genes that do not occur in any known pathways were removed according to the Saccharomyces Genome Database (SGD) (26) which contains 142 known Yeast pathways covering 515 genes. There were 502 genes finally left in the three expression data sets, which were subsequently used as input in our experiments. We note that each of these 502 genes participates into at least one pathway, and some of them participate into up to seven different pathways. The number of genes involved in these 142 pathways varies from 1 to 23 with 5 on average.

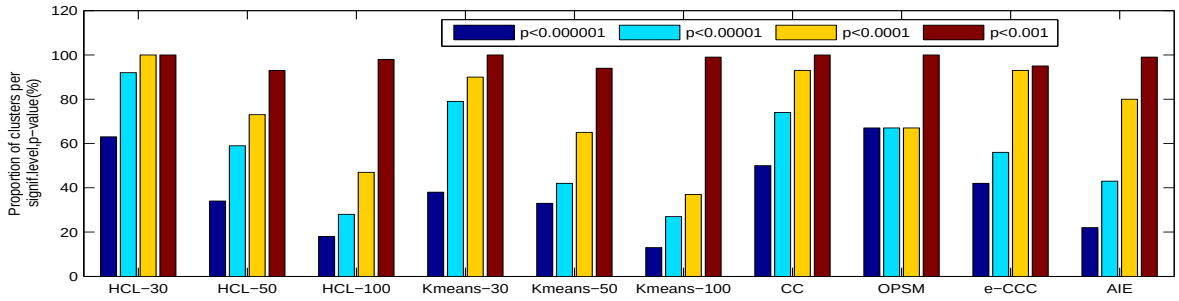
We compare the P-values (biological significance) and R-values (negative significance) of AIE patterns with those of the gene clusters found by the widely used clustering/biclustering methods. The purpose of this comparison is to confirm that many important negative correlations were unable to be identified by those conventional algorithms. We also present representative AIE patterns, and other statistics information of AIE patterns, including the up-down index values and the differential gap information. More importantly, we take case studies to illustrate how the AIE patterns are biologically interpreted for enhancing the analysis of molecular pathways.

Comparison by using P-values and R-values

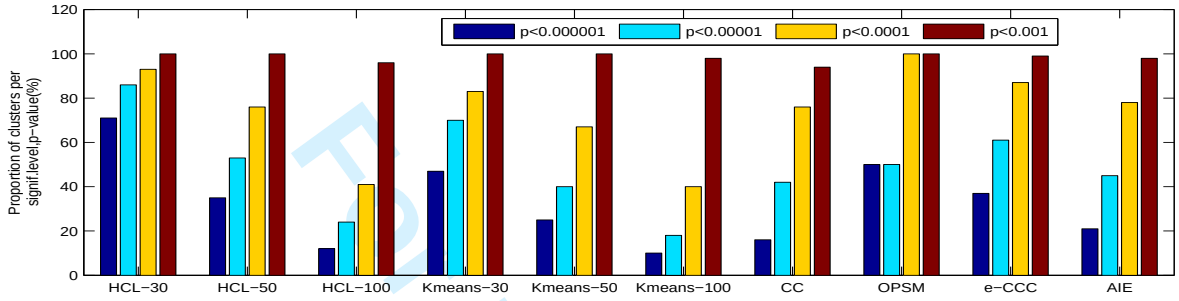
Conventional clustering methods under our comparison include a hierarchical clustering method (HCL) (19, 27), a K-means clustering method (Kmeans) (27), the Cheng&Church biclustering method (CC)(3, 27), the order preserving sub-matrix algorithm (OPSM) (3, 27) and e-CCC-Biclustering (e-CCC) (23). All of their implementation are available at BicAT (27) or at BiGGEsTS (28).

It's specially noted that the notion of CCC-Biclusters (22) and its extension e-CCC-Biclusters are closely related to our concept of AIE patterns, in particular when the sign-change rule (U $\leftrightarrow$ D) is combined to form CCC-Biclusters or e-CCC-Biclusters. By definition, a CCC-Bicluster has to satisfy the condition that every possible gene pair in this bicluster shares a positive (coherent) expression change behavior over the time. e-CCC-Bicluster extends CCC-Bicluster by allowing a certain degree of noise (measured by the parameter e) in a CCC-Bicluster, such that the coherent expression change behavior in an e-CCC-Bicluster may not be always the same on some time points. The sign-change rule introduced in the e-CCC-Bicluster mining algorithm (23) enriches the diversity of these biclusters, and it can be used to detect the so called anti-correlated patterns.

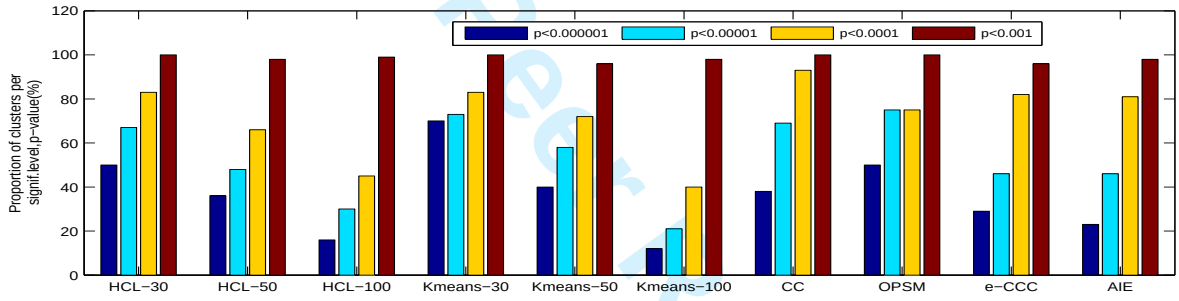
As introduced, the definition of AIE patterns is simple and different. Two non-overlapping gene groups I_1 and I_2 can form an AIE pattern if and only if I_1 and I_2 have a negative expression change behavior over a time segment (Definition 1). That is, the genes in the same group (I_1 or I_2) are not necessarily required to have exact coherent expression change over the time segment. Even if the *sign-change* rule is applied to AIE patterns, the gene pairs within $I_1 \cup I_2$ may still not have much coherence. Though an anti-correlated pattern may sometimes become an AIE pattern, on the other hand, an AIE pattern usually does not satisfy the conditions required by an anti-correlated pattern. Therefore, they two are not equivalent. Another difference lies in the algorithms of mining e-CCC-Bicluster and AIE patterns. Although both



(a) Proportion of P-values in *Alpha factor*



(b) Proportion of P-values in *Cdc15*



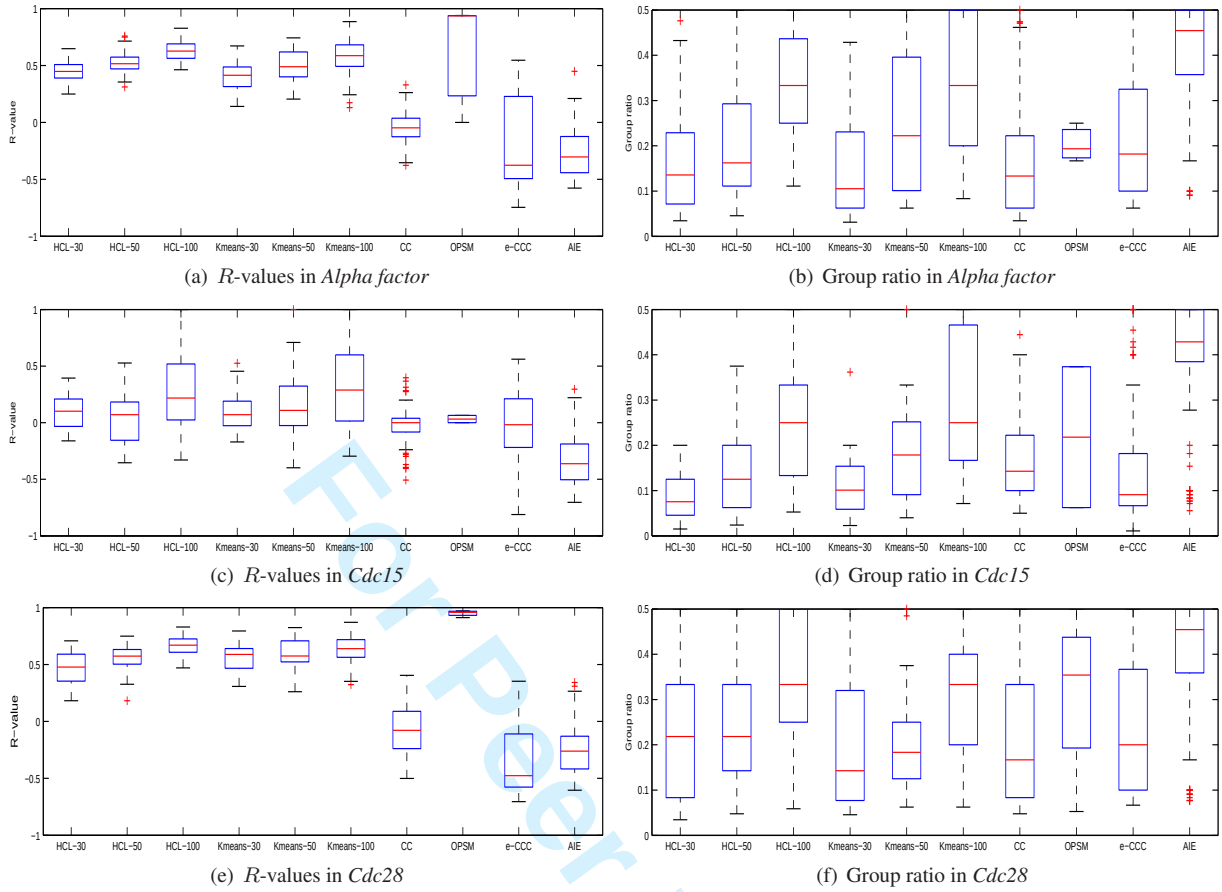
(c) Proportion of P-values in *Cdc28*

Figure 4. Biological significance comparison between our AIE patterns and gene clusters by the conventional methods. Here, HCL-*x* or Kmeans-*x* stands for *x* number of clusters being pre-set.

of them take an exhaustive enumeration approach, e-CCC-Biclustering needs the parameter e to control the noise level of e-CCC-Biclusters and also needs sign-change to allow negative correlation, while AIE pattern mining is a parameter-free algorithm (many proposed indexes such as up-down index and differential gap are just used in the post-analysis, although these indexes can also be used as predefined parameters to reduce the running time of AIE pattern mining). We would like to also point out that both CCC-Biclusters and AIE patterns use the maximality rule to get rid of some redundancy in the patterns.

In our comparison, the input parameters of these conventional methods are all set as the default values as suggested in (23) and (27). In order to avoid possible comparison bias, some of the gene clusters obtained by the biclustering methods are filtered if the overlap degree between any two clusters is more than 25% as previously done by (22, 23, 27).

We first compare P-values (29) which can indicate a biological significance of a gene cluster indirectly. In this paper, P-values are calculated through gene set enrichment analysis with **Fisher's exact test** (29). A P-value less than 10^{-3} is widely accepted as a gold standard in most biological significance analysis. Fig. 4 shows the bar charts representing proportions of gene clusters whose P-values are less than some thresholds. For example, from the *Cdc28* data set, there are 57592 AIE patterns identified. 151 of them are left after the filtering, in which 99% have a P-value less than 10^{-3}, and 22% with a P-value less than 10^{-6}. In many cases, AIE patterns have better proportions of biologically significant gene clusters than those of the gene clusters found by the conventional methods. It is worth noting that the OPSM method seems to perform better than the other biclustering methods. A possible reason is that it only outputs no more than 30 *top* clusters and no more than 5 clusters after filtering (shown in the second column of Table 1). CC and e-CCC-Biclustering are slightly better than our AIE method in terms of these P-values. We also note that when the number of clusters is set to be big by HCL or Kmeans, the genes contained in each cluster tend to be small. As the size of a cluster affects the calculation of P-values, the performance of HCL or Kmeans is influenced greatly by the pre-given number of clusters. See Table 1 for detailed information of the clusters under different settings.



(a) R-values in *Alpha factor* (b) Group ratio in *Alpha factor*

(c) R-values in *Cdc15* (d) Group ratio in *Cdc15*

(e) R-values in *Cdc28* (f) Group ratio in *Cdc28*

Figure 5. The generalized Pearson's correlation coefficient and group ratio comparison between AIE pattern mining and the conventional methods.

We have seen that AIE patterns have close competitive P-values with those of the gene clusters generated by the conventional clustering methods. Next, we report the R-values and group ratio information of AIE patterns. Fig. 5(a), Fig. 5(c) and Fig. 5(e) show a box-plot view of the R-values of the gene clusters from the *Alpha factor*, *Cdc15*, and *Cdc28* data set, respectively, by different methods. We can see that the R-values of the AIE patterns are almost all less than 0. This result is in full agreement with our original notion of negative correlation. In particular, nearly half of the AIE patterns have a R-value lower than -0.5. However, among the gene clusters found by the traditional clustering methods and OPSM, usually only a small proportion of them have negative R-values. So combining the R-values and P-values, we can note that only CC, e-CCC-Biclustering, and our AIE approach are good to detect biologically significant *and* negative correlations. We also consider the group ratio information $GR(I_1,I_2)$ of the clusters in the comparison. The CC and e-CCC-Biclustering method both tend to find clusters with small group ratio values around 0.2, while our AIE patterns usually have a group ratio value higher than 0.4. This indicates that CC and e-CCC-Biclustering prefer negative correlations between unbalanced gene group pairs. However, our AIE approach can indeed find negative correlation spanning two size-balanced gene groups which can have strong biological significance. See Fig. 5(b), Fig. 5(d) and Fig. 5(f)) for a detailed comparison of the group ratio information.

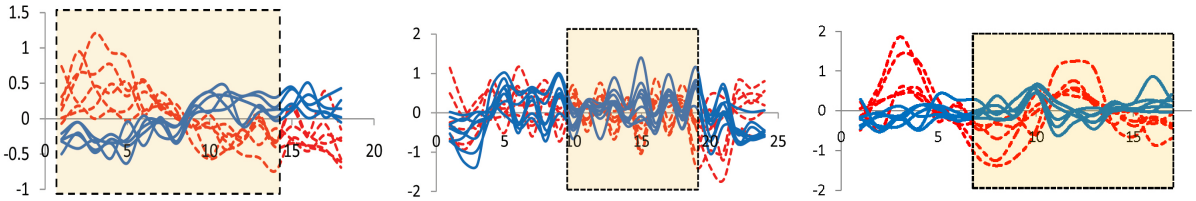
Representative AIE patterns from the three expression data sets

Table 2 presents statistics information of all AIE patterns discovered from the three data sets when the gene number threshold for $|I_1|$ and $|I_2|$ is set as 5, and the minimal number of time points in $|J|$ is set as 10. In this table, the column "**# of AIE Patterns**" refers to the total number of AIE patterns from one data set, together with their average number of genes and their average time points in one pattern; the column **Differential gap** indicates the smallest, biggest, and average expression differences between the two subgroups for the AIE patterns in each data set; the column **Up-down index** indicates the smallest, biggest, and average up-down index values for the AIE patterns in each data set; and, the column **P-Value** shows the minimal, maximal, and average P-values of the AIE patterns in each data set. From this table, we can see that the shape and property of AIE patterns can vary very much. There are many other choices to set the size threshold for $|I_1|$, $|I_2|$, and $|J|$; readers are referred to use our website http://sunim1.birc.ntu.edu.sg/~aie/ to get the statistics information when a different threshold is set.

Fig. 6 displays three typical examples of negative correlation under the three experiment conditions. These negative correlations can be categorized into two kinds of

Table 2. Some statistics information on the complete sets of AIE patterns.

Data sets	# of AIE patterns & (avg. I, avg. J)	Differential gap (min, max, avg.)	Up-down index (min, max, avg.)	P-value (min, max, avg.)
Alpha factor	8508 (12, 10)	(0.013, 0.311, 0.075)	(0.0, 1.0, 0.20)	(1.36E-13, 0.002, 3.34E-5)
Cdc15	279473 (19, 11)	(0.007, 0.312, 0.073)	(0.0, 1.0, 0.80)	(0.0, 0.003, 1.53E-5)
Cdc28	57592 (15, 11)	(0.010, 0.328, 0.086)	(0.0, 0.67, 0.157)	(0.0, 0.002, 2.11E-5)



(a) From Alpha factor (b) From Cdc15 (c) From Cdc28

Figure 6. Representative examples of AIE pattern from the three data sets. Here, two groups of genes colored in red and blue, show negative correlation during the time points in orange area.

behavior according to their up-down index values. Under the experiment condition *Alpha factor* or *Cdc28*, the expressions of the two gene groups do not up-down change frequently. Fig. 6(a) shows such an expression trend where the expression up-down change happened only once which was at the time point 9.

However, most AIE patterns under the condition *Cdc15* can have up-down index values close to 1.0. See the fourth column of Table 2. This means that the gene groups change their expression up-down very frequently under this cell environment. Fig. 6(b) shows a perfect example for this expression trend that continuously crosses 14 time points. Therefore, we can conjecture that negative correlation can behavior in tremendously different ways in the Yeast cell cycle when different conditions are applied, meanwhile their P-values are all very significant (as shown in the 5th column of Table 2 as well as in Fig. 4).

Biological interpretation of AIE patterns: three case studies

We take special case studies to illustrate how AIE patterns and negative correlations can be used to infer new modules of gene regulation, transcription factors (TFs), and regulatory networks. Our first example is the representative AIE pattern identified from the *Alpha factor* data set, whose gene expression profile is shown in Fig. 6(a). Total 11 genes are involved in the two gene subgroups of this pattern that have a negative correlation spanning 14 continuous time points. One sub-group consists of 6 genes; We specially name it "red" group and denote by $I_r = \{YIR038C$ (GeneID:854856), $YNR001C$ (GeneID:855732), $YKL127W$ (GeneID:853732), $YOL126C$ (GeneID:853994), $YJL068C$ (GeneID:853377), $YLR328W$ (GeneID:851039)$\}$. The other group consists of 5 genes; We specially name it "blue" group and denote by $I_b = \{YDR234W$ (GeneID:851820), $YDL182W$ (GeneID:851346), $YDL131W$ (GeneID:851425), $YBR265W$ (GeneID:852568), $YNR050C$ (GeneID:855786)$\}$. We also denote this pattern simply as AIE-*Alpha*-87. It is one of the most negative AIE patterns in *Alpha factor* according to their R-values; and its P-value is 2.9×10^{-10}. The main covering pathway of AIE-*Alpha*-87 is *lysine biosynthesis*, which has 7 genes known currently and 4 of them are contained in the blue group I_b.

Table 3. Possible *YEL009C* binding sites at the upstream of the 5 genes in the blue group of AIE-*Alpha*-87

Object Genes	Possible binding sites
YDL131W	TGACTGA, TTGCGCAA
YDL182W	TGACTGA
YDR234W	TGACTGA
YNR050C	TGACTGA, TGACTMT
YBR265W	TGACTMT

A new regulatory module and its putative transcription factor Let's start the analysis on the five genes in the blue group I_b of AIE-*Alpha*-87. As mentioned, four genes in I_b are directly involved in the *lysine biosynthesis* pathway. In fact, the four genes are co-regulated by a known TF $YEL009C$. As all of the five genes in I_b are inherently co-expressed, we can infer that they altogether form a regulatory module with $YEL009C$ as a co-TF. To confirm this hypothesis,

we examined the whole upstream of each gene to identify their binding motifs by using YEASTRACT (30). Three possible $YEL009C$ binding sites were identified (Table 3). We can infer that the first four genes in this table share a binding motif "TGACTGA", while the last two genes $YNR050C$ and $YBR265W$ share a binding motif "TGACTMT". Thus, these five genes are all likely to be co-regulated by $YEL009C$ through its binding upon regulation segments in the upstream of the five genes. This is a new insight into the gene regulatory behavior of this module and its transcription factor. This new understanding is mainly attributed to the positive relationship of the genes within the blue group.

Building a tree-structure regulatory network On top of the idea of Boolean regulatory networks (31), we incorporate our negative correlations and introduce a *tree-structure* regulatory network for genes involved in an AIE pattern. We take two steps to complete the induction of these trees. The first step is to use YEASTRACT (30) to construct an

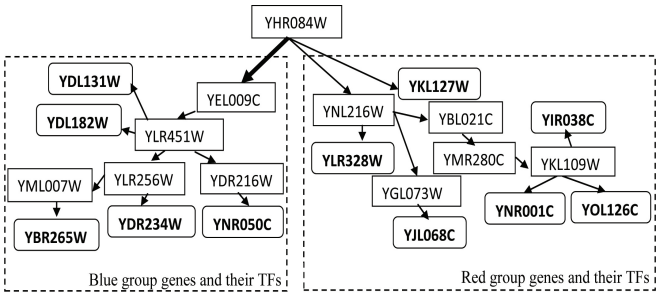


Figure 7. A sub-regulatory tree for AIE-*Alpha*-87.

initial regulatory network (RN) by taking as input all the genes in an AIE pattern as well as all of their known TFs in Yeast; the second step is to trim the RN to eventually become a tree structure. The trimming constraints are set as follows: (i) all the leaf nodes of the tree are required to represent the genes in the AIE pattern, (ii) the inner nodes all represent the known TFs, and (iii) the tree can be decomposed into two sub-trees by removing one edge in such a way that each sub-tree exactly covers all genes in one group of the AIE pattern. This edge under the removal is critically important. The regulation represented by this edge should be either 'enabled' or 'disabled', and there should be at least one negative regulator to produce this negative correlation between the expressions of the two gene groups. Fig. 7 shows a tree-structure regulatory network built from AIE-*Alpha*-87.

This regulatory sub-tree possesses strong biological meanings. First of all, the regulatory relations indicated as edges in this tree all have direct/indirect evidence supported by known literatures (30). In particular, $YLR451W$ is a negative regulator as reported and studied in (32), while other TFs are positive regulators. Next, we explain why genes in the blue group can show the negative correlation with the genes in the red group. This is likely due to the participation of the negative regulator $YLR451W$ (32), which is the next TF after $YEL009C$ on the regulatory paths from the root $YHR084W$ to the genes in the blue group. As the TFs on the paths from the root to the red genes are completely different from the TFs on the paths from the root to the blue genes, it is most likely that the negative co-regulation between the genes in the two groups is caused by $YHR084W$ with an auxiliary help from $YLR451W$. In fact, a recent study found that $YHR084W$ is a specific Yeast cell cycle transcription factor (33). Thus, we can see that an AIE pattern can not only identify two groups of genes with a negative correlation on gene expression profiles, and can also imply that one group genes possibly come from a same pathway, having similar expression behavior as reference to the genes in the other group. Therefore, such negative correlations can uncover the connections among TFs which are at a higher layer in the regulatory network.

New pathway components Through the above analysis, we have already understood that there exists a possible regulation pathway of $YEL009C$ on the five genes in the blue group, and that the six genes in the red group have a negative correlation with *lysine biosynthesis* which is likely caused by the co-regulation of $YHR084W$ and the negative regulation

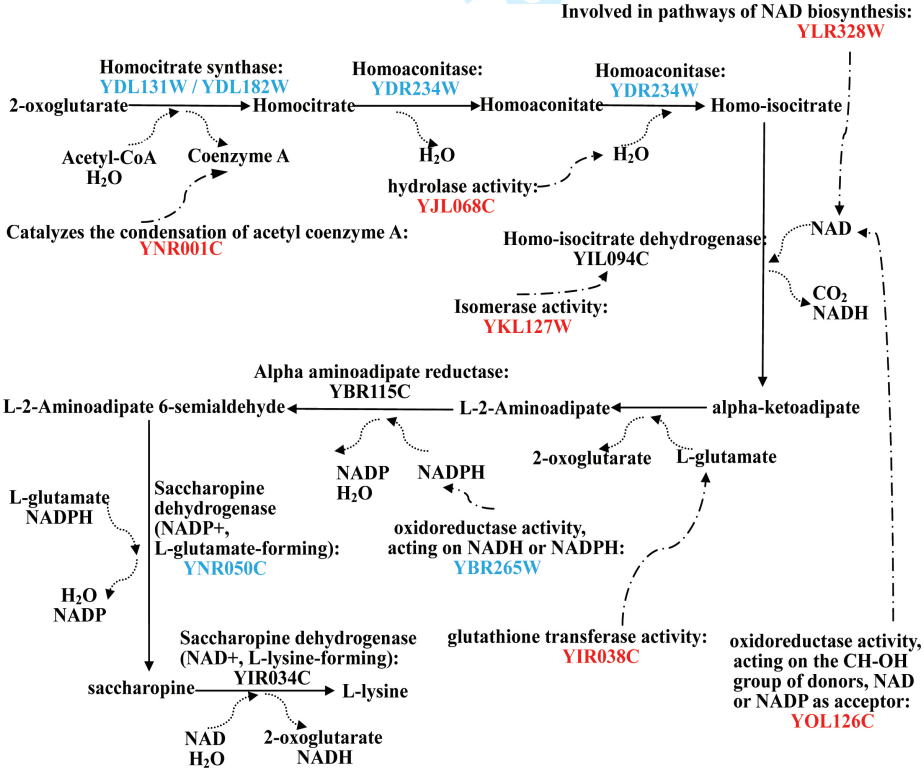


Figure 8. An expanded diagram for the pathway *lysine biosynthesis* after our functional annotation is used to derive new components (shown as dashed lines) based on the negative correlation of AIE-*Alpha*-87.

of $YLR451W$. We next study whether the 11 genes of AIE-*Alpha*-87 have any biological function relations with lysine biosynthesis.

A known *lysine biosynthesis* pathway diagram can be obtained from SGD (26). As mentioned, four ($YDL131W$, $YDL182W$, $YDR234W$, and $YNR050C$) of the 11 genes of the AIE-*Alpha*-87 pattern participate directly in this pathway. To see whether the other 7 genes have any function related to a biological molecular or a biochemistry reaction in the current lysine biosynthesis, we attempted a function annotation. A more complete diagram for the *lysine biosynthesis* pathway is shown in Fig. 8, where the dashed line means a new component assessed by the annotation. In detail, genes $YBR265W$ and $YOL126C$ can affect on molecules NAD and NADPH through oxidation. Genes $YNR001C$, $YKL127W$, $YIR038C$, and $YJL068C$ are also indirectly related to this pathway through their biochemical functions on the direct partners of *lysine biosynthesis*. The most amazing gene is $YLR328W$. This gene is a key component in the pathway of *NAD biosynthesis* which produces NAD, an important molecule for the lysine biosynthesis. Therefore, all the 11 genes have a direct or indirect relationship with the *lysine biosynthesis* pathway. If taking only positive correlation for the study, many genes indirectly related to the pathway like those in the red group of AIE-*Alpha*-87 would be ignored. Therefore, through function analysis on AIE patterns, we can expand existing pathways by adding indirect biochemistry reactions or by linking to other biosynthesis pathways. This is the main reason why we say AIE pattern can enhance our understanding on molecular pathways.

The second example of our case study is to show how a negative correlation of the genes in the same pathway can be used to identify gene targets in biological experiments for testing the negative mechanism.

Gene target identification for testing negative mechanism We found that many of our AIE patterns consist of two groups of genes that are from the same pathways. One example is the representative AIE pattern in *Cdc15* condition named AIE-*Cdc15*-273421, whose gene expression profile is shown in Fig. 6(b). Its P-value is $4.54E-12$, R-value is about -0.55, up-down index is 1.0, and differential gap is about 0.068. A reason to choose this pattern for illustration is because of its extremely high up-down change frequency (i.e. 1.0). Of the total 14 genes in this pattern, there are six genes contained in the pathway *ergosterol biosynthesis*. Three of them ($YML008C$ (GeneID:855003), $YGL001C$ (GeneID:852883) and $YLR100W$ (GeneID:850790)) are located at upstream, while the rest three ($YMR202W$ (GeneID:855242), $YGL012W$ (GeneID:852872) and $YMR015C$ (GeneID:855029)) are at the downstream. According to a known structure of the pathway *ergosterol biosynthesis* partially shown in Fig. 9, these 6 genes function sequentially one by one to achieve the final function of this pathway. Associating these genes' expressions with their roles in this pathway, it can be observed that when upstream genes are up-regulated or down-regulated, the genes located at the downstream have simultaneous opposite expressions. Intuitively, we can infer that there exists some negative control mechanism in the path from fecosterol to episterol, which is the functional boundary between the upstream and downstream gene groups (Fig. 9). Interestingly, it has

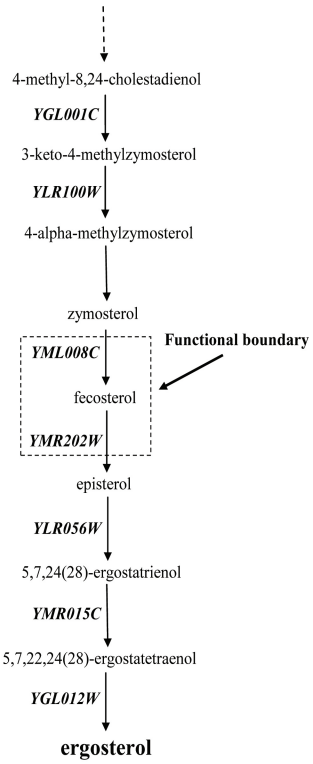
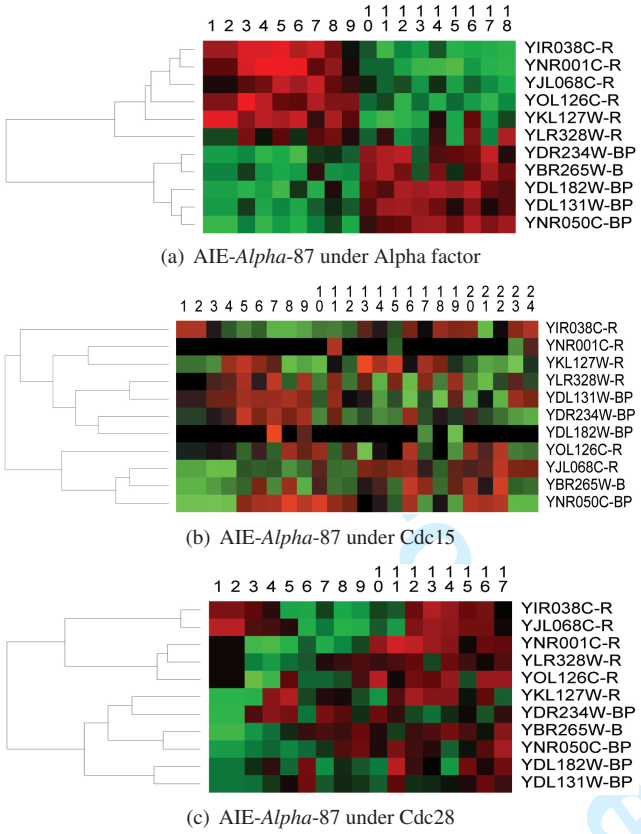


Figure 9. A partial diagram of the pathway *ergosterol biosynthesis*.

been previously reported that ergosterol exerts a negative feedback on its own biosynthesis in *S.cerevisiae*, particularly at the C-24 methylation step involving the gene $YML008C$ (34, 35). Therefore, we can see that the negative correlation of the genes in the same pathway is potentially useful to identify gene targets in the biological experiments for testing negative mechanisms.

Invariable negative correlations under different conditions We present our third case study and examine hierarchical clusterings of the expression profiles of the genes in one AIE pattern under the three different cell cycle environments. Taking again AIE-Alpha-87 as an example, hierarchical clusterings of the expression profiles of the 11 genes in *Alpha factor*, *Cdc15*, and *Cdc28* are shown in Fig. 10 (The drawing was done by the software PermutMatrix (36).) In the case of *Alpha factor*, the 11 genes are nicely divided into two groups: one matches with the blue group, the other maps to the red group. However, the negative correlation under *Alpha factor*, disappeared in the other two environments of cell cycle. We can also see that the 4 known genes in the pathway *lysine biosynthesis* always have co-behaviors under different conditions of cell cycle, but the other genes' negative correlation cannot be always maintained.

Thus, another perspective to understand AIE patterns is to see whether one pattern as a whole can be conserved during different biological environments. Of course, the above example is not the case. In fact, there are very little overlapping between any two AIE pattern pairs from different environments of cell cycle. This is because most environmental perturbations can cause big change in expression, resulting in alteration in the complex regulatory

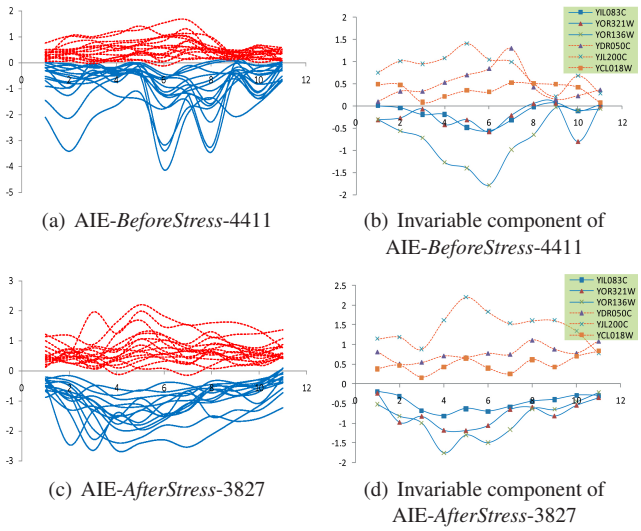


(a) AIE-*Alpha*-87 under Alpha factor

(b) AIE-*Alpha*-87 under Cdc15

(c) AIE-*Alpha*-87 under Cdc28

Figure 10. The expression profiles of the genes in AIE-*Alpha*-87 under 3 different conditions (*Alpha factor*, *Cdc15* and *Cdc28*). The signs following each gene name have specific meanings. **-B** stands for this gene in the blue group; **-BP** stands for this gene in the blue group and also belonging to the pathway; and **-R** stands for this gene in the red group.



(a) AIE-*BeforeStress*-4411

(b) Invariable component of AIE-*BeforeStress*-4411

(c) AIE-*AfterStress*-3827

(d) Invariable component of AIE-*AfterStress*-3827

Figure 11. An invariable negative correlation shared by six common genes of two AIE patterns (AIE-*BeforeStress*-4411 and AIE-*AfterStress*-3827).

system. Ronen *et al.* had designed a glucose plus experiment to detect affection of the small constraint perturbation (37), which is on the other hand helpful to understand stable negative correlations. We briefly explain two AIE patterns identified from their data set. One AIE pattern before glucose plus (denoted as AIE-*StressBefore*-4411) contains 31 genes whose expression profiles are shown at Fig. 11(a). An AIE pattern after glucose plus (denoted as AIE-*StressAfter*-3827) contains 30 genes whose expression profiles are shown at Fig. 11(c). Interestingly, these two patterns share six genes that maintain the same negative correlation. The six genes are: $YDR050C$ (GeneID:851620), $YJL200C$ (GeneID:853230), $YCL018W$ (GeneID:850342) and $YIL083C$ (GeneID:854726), $YOR321W$ (GeneID:854499), $YOR136W$ (GeneID:854303). The expression profiles of this stable negative correlation are displayed at Fig. 11(b) and Fig. 11(d). It looks that these six genes and their negative correlation are a stable genetic indicator during glucose plus. From the viewpoint of gene function, the other 25 genes in AIE-*StressBefore*-4411 have a significant function hit on *transferase activity* (10 out of 25 genes), while the other 24 genes in AIE-*StressAfter*-3827 have a significant function hit on *oxidoreductase activity* (10 out of 24 genes). It is also known that oxidoreductase is closely related to high-glucose ambience (38). Therefore, it is suggestive that this stable AIE pattern shared by these six genes is likely involved in both normal and stress activity of Yeast.

CONCLUSION

The main contribution of this work is the formalization of the widely observed negative correlations in genes' functions within molecular pathways. Through our mining algorithm which uses a suffix-tree data structure and a bi-clique search idea, all possible AIE patterns in a time-course gene expression data set can be enumerated. As some of them are perhaps of less interests, we have suggested to use the size threshold, up-down index, and R-value index to control the quantity and quality of AIE patterns in the post-analysis. Although pairs of gene clusters computed by the traditional clustering methods can find some negative correlations, they are unable to detect negative correlations shown in time segments as our AIE patterns can do. The biclustering methods can iteratively conduct clustering from both genes and time points, it is still hard to detect all negative correlation candidates in large data set. However, our mining algorithm can overcome this difficulty.

Our experimental results on three Yeast cell cycle expression data sets have demonstrated that maximal negative correlation can occur between pairs of large groups of genes, one group or both covering many genes from the same pathway. Basing on existing knowledge about molecular pathways, negative correlations can be used to infer new gene regulation modules, transcription factors, and regulatory networks, as shown in our case studies. With these new elements, we are able to get a fuller and deeper picture about the direct and/or indirect relationships of all components in a molecular pathway. Besides, significant invariable negative correlations are found in both normal activity and stress activity of Yeast. All these ideas and results highlight that maximal negative correlation is an important characteristic in the gene expression profiles within pathways, which is expected to be useful in the cell stress response study, drug

response evaluation, and cancer related pathways' detection, etc.

ACKNOWLEDGEMENTS

This research work was funded by a Tier-1 grant (RG66/07) awarded by Nanyang Technological University, Singapore.

Conflict of interest statement. None declared.

REFERENCES

1. Segal, E., Wang, H., and Koller, D. (2003) Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics,* **19 suppl 1**, i264–i271.
2. Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., , and Davis, R. W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell,* **2**(1), 65–73.
3. Madeira, S. C. and Oliveira, A. L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* **1**(1), 24–45.
4. Aguilar-Ruiz, J. S. (2005) Shifting and scaling patterns from gene expression data. *Bioinformatics,* **21**(20), 3840–3845.
5. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics,* **34**(2), 166–176.
6. Cheng, K. O., Law, N. F., Siu, W. C., and Liew, A. W. (2008) Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization. *BMC Bioinformatics,* **9**(210).
7. Kim, J. and Kim, H. (2008) Clustering of change patterns using fourier coefficients. *Bioinformatics,* **24**(2), 184–191.
8. Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., and Lohmann, J. U. (2005) A gene expression map of arabidopsis thaliana development. *Nature Genetics,* **37**(5), 501–506.
9. James, C. G., Appleton, C. T., Ulici, V., Underhill, T. M., and Beier, F. (2005) Microarray analyses of gene expression during chondrocyte differentiation identifies novel regulators of hypertrophy. *Molecular Biology of the Cell,* **16**(11), 5316–5333.
10. Chuang, C. L., Jen, C. H., Chen, C. M., and Shieh, G. S. (2008) A pattern recognition approach to infer time-lagged genetic interactions. *Bioinformatics,* **24**(9), 1183–1190.
11. Stekel, D. J. and Jenkins, D. J. (2008) Strong negative self regulation of prokaryotic transcription factors increases the intrinsic noise of protein expression. *BMC Systems Biology,* **2**(6).
12. Missero, C., Pirro, M. T., and Di Lauro, R. (2000) Multiple ras downstream pathways mediate functional repression of the homeobox gene product ttf-1. *Molecular and Cellular Biology,* **20**(8), 2783–2793.
13. Segal, L., Lapidot, M., Solan, Z., Ruppin, E., Pilpel, Y., and Horn, D. (2007) Nucleotide variation of regulatory motifs may lead to distinct expression patterns. *Bioinformatics,* **23**(13), i440–i449.
14. Shieh, G. S., Chen, C. M., Yu, C. Y., Huang, J., Wang, W. F., and Lo, Y. C. (2008) Inferring transcriptional compensation interactions in yeast via stepwise structure equation modeling. *BMC Bioinformatics,* **9**(134).
15. Millar, C. B. and Grunstein, M. (2006) Genome-wide patterns of histone modifications in yeast. *Nature Reviews Molecular Cell Biology,* **7**(9), 657–666.
16. Esteller, M. (2007) Cancer epigenomics: Dna methylomes and histone-modification maps. *Nature Reviews Genetics,* **8**(4), 286–298.
17. Wiencke, J. K., Zheng, S., Morrison, Z., and Yeh, R. F. (2008) Differentially expressed genes are marked by histone 3 lysine 9 trimethylation in human cancer cells. *Oncogene,* **27**(17), 2412–2421.
18. McGarvey, K. M., Van Neste, L., Cope, L., Ohm, J. E., Herman, J. G., Van Criekinge, W., Schuebel, K. E., and Baylin, S. B. (2008) Defining a chromatin pattern that characterizes dna-hypermethylated genes in colon cancer cells. *Cancer Research,* **68**(14), 5753–5759.
19. Yuan, Y., Li, C. T., and Wilson, R. (2008) Partial mixture model for tight clustering of gene expression time-course. *BMC Bioinformatics,* **9**(287).
20. Ji, L. and Tan, K. L. (2005) Identifying time-lagged gene clusters using gene expression data. *Bioinformatics,* **21**(4), 509–516.
21. Supper, J., Strauch, M., Wanke, D., Harter, K., and Zell, A. (2007) Edisa: extracting biclusters from multiple time-series of gene expression profiles. *BMC Bioinformatics,* **8**(334).
22. Madeira, S. C., Teixeira, M. C., Sá-Correia, I., and Oliveira, A. L. (2008) Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* **http://doi.ieeecomputersociety.org/10.1109/TCBB.2008.34**.
23. Madeira, S. C. and Oliveira, A. L. (2009) A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series. *Algorithms for Molecular Biology,* **4**(8).
24. Li, J., Liu, G., Li, H., and Wong, L. (2007) Maximal biclique subgraphs and closed pattern pairs of the adjacency matrix: A one-to-one correspondence and mining algorithms. *IEEE Transactions on Knowledge and Data Engineering,* **19**(12), 1625–1637.
25. 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 microarray hybridization. *Molecular Biology of the Cell,* **9**(12), 3273–3297.
26. Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998) Sgd: Saccharomyces genome database. *Nucleic Acids Research,* **26**(1), 73–79.
27. Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics,* **22**(9), 1122–1129.
28. Gonalves, J. P., Madeira, S. C., and Oliveira, A. L. (2009) Biggests: integrated environment for biclustering analysis of time series gene expression data. *BMC Research Notes,* **2**(124).
29. Curtis, R. K., Oresic, M., and Vidal-Puig, A. (2005) Pathways to the analysis of microarray data. *Trends in Biotechnology,* **23**(8), 429–435.
30. Monteiro, P. T., Mendes, N. D., Teixeira, M. C., d'Orey, S., Tenreiro, S., Mira, N. P., Pais, H., Francisco, A. P., Carvalho, A. M., Lourenço, A. B., Sá-Correia, I., Oliveira, A. L., and Freitas, A. T. (2008) Yeastract-discoverer: new tools to improve the analysis of transcriptional regulatory associations in saccharomyces cerevisiae. *Nucleic Acids Research,* **36**(Database issue), D132–D136.
31. Schlitt, T. and Brazma, A. (2007) Current approaches to gene regulatory network modelling. *BMC Bioinformatics,* **8 Suppl 6**(S9).
32. Guelzim, N., Bottani, S., Bourgine, P., and Képès, F. (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics,* **31**(1), 60–63.
33. Wu, W. S. and Li, W. H. (2008) Systematic identification of yeast cell cycle transcription factors using multiple data sources. *BMC Bioinformatics,* **9**(522).
34. Veen, M., Stahl, U., and Lang, C. (2003) Combined overexpression of genes of the ergosterol biosynthetic pathway leads to accumulation of sterols in saccharomyces cerevisiae. *FEMS Yeast Research,* **4**(1), 87–95.
35. Vandeputte, P., Tronchin, G., Larcher, G., Ernoult, E., Bergès, T., Chabasse, D., and Bouchara, J. P. (2008) A nonsense mutation in the erg6 gene leads to reduced susceptibility to polyenes in a clinical isolate of candida glabrata. *Antimicrobial Agents and Chemotherapy,* **52**(10), 3701–3709.
36. Caraux, G. and Pinloche, S. (2005) Permutmatrix: a graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics,* **21**(7), 1280–1281.
37. Ronen, M. and Botstein, D. (2006) Transcriptional response of steady-state yeast cultures to transient perturbations in carbon sourse. *Proc. Natl. Acad. Sci. USA,* **103**(2), 389–394.
38. Nayak, B., Xie, P., Akagi, S., Yang, Q., Sun, L., Wada, J., Thakur, A., Danesh, F. R., Chugh, S. S., and Kanwar, Y. S. (2005) Modulation of renal-specific oxidoreductase/myo-inositol oxygenase by high-glucose ambience. *Proc. Natl. Acad. Sci. USA,* **102**(50), 17952–17957.

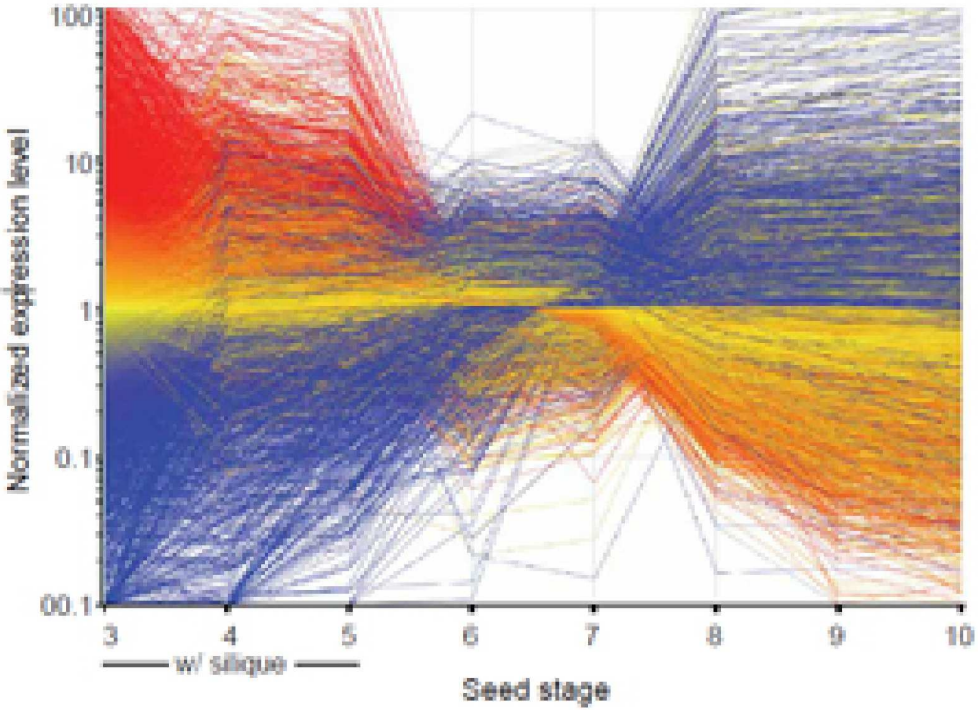


Figure 1.(a) Opposing expression
202x143mm (600 x 600 DPI)

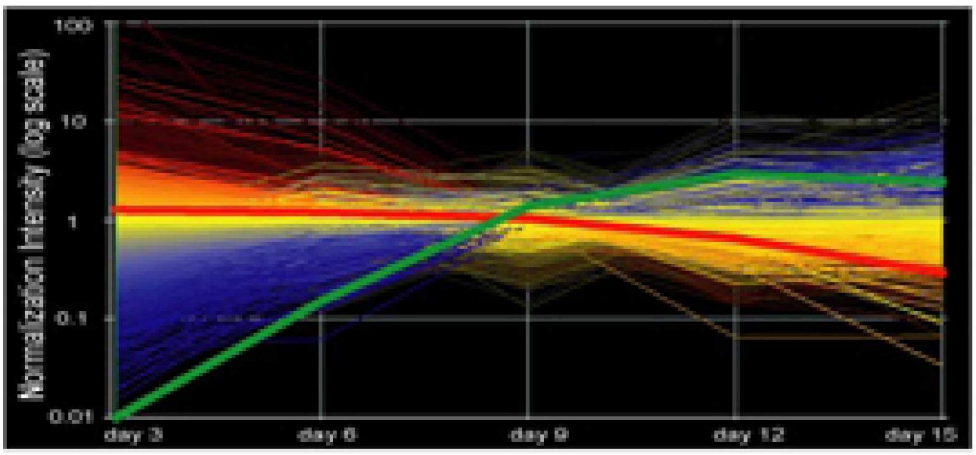


Figure 1.(b) Opposite expression
202x96mm (600 x 600 DPI)

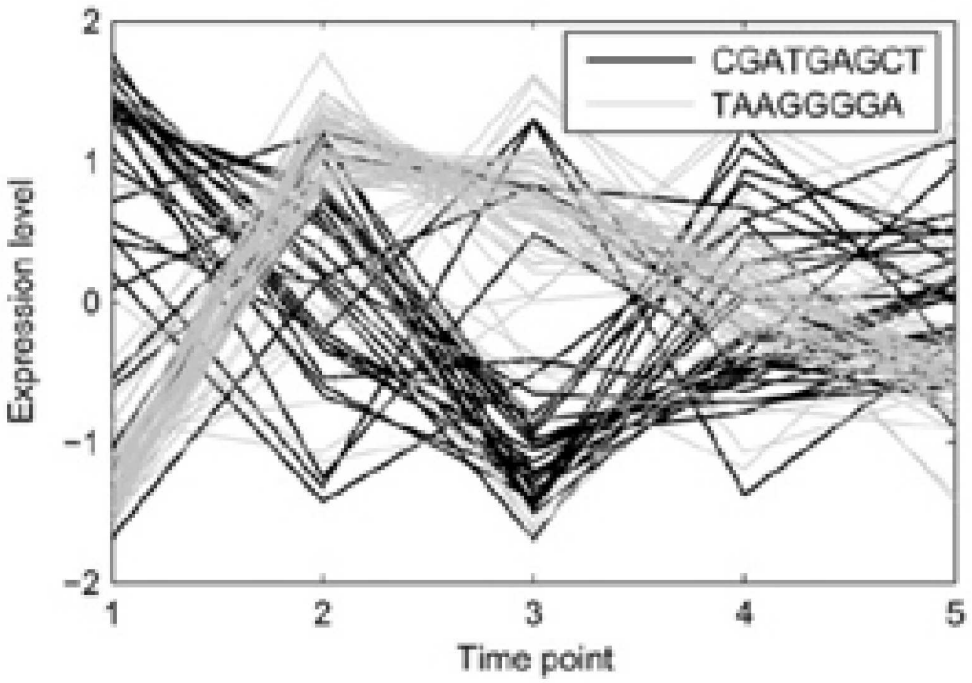


Figure 1.(c) Contrasting expression of regulatory motifs
202x140mm (600 x 600 DPI)

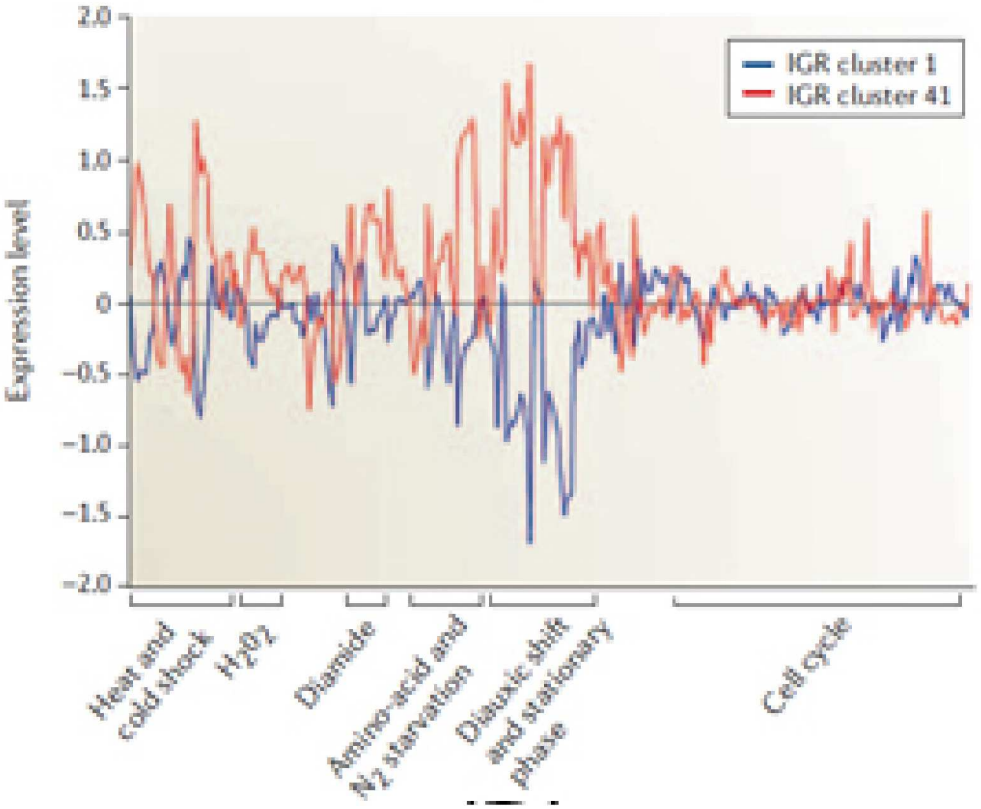


Figure 1.(d) Expression of compensatory genes
201x162mm (600 x 600 DPI)

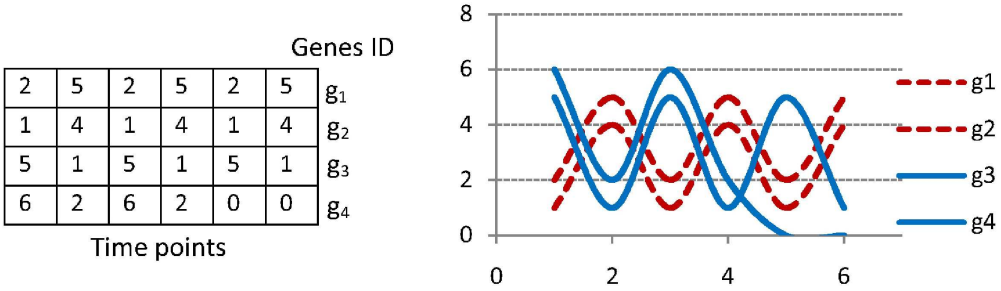


Figure 2.(a) An original gene expression data matrix
197x56mm (600 x 600 DPI)

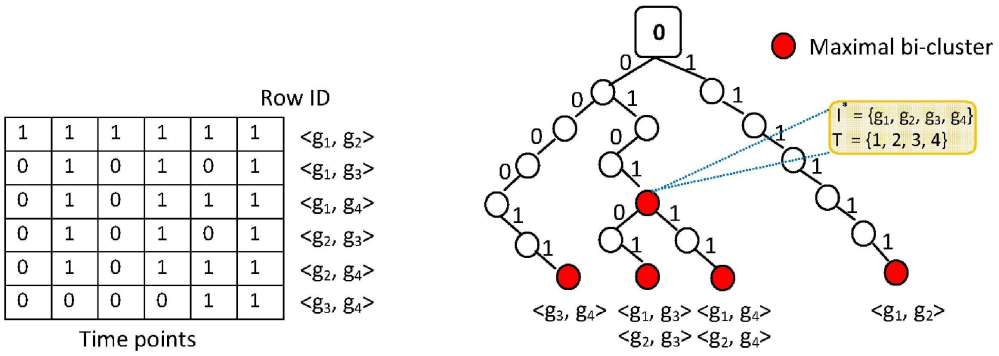


Figure 2.(b) The dichotomy matrix and its suffix tree
196x68mm (600 x 600 DPI)

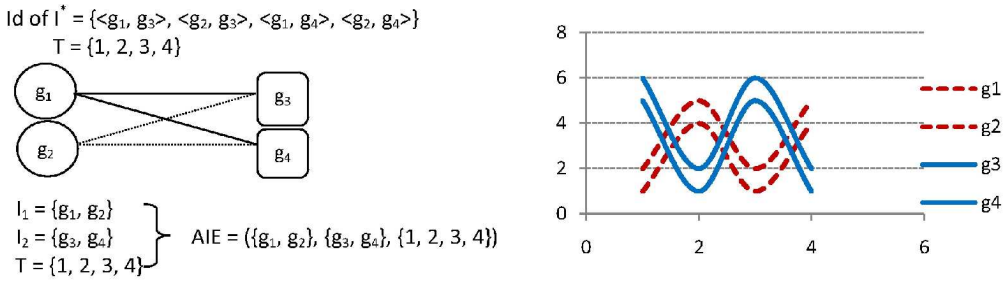


Figure 2.(c) One bi-clique and its corresponding AIE pattern
199x56mm (600 x 600 DPI)

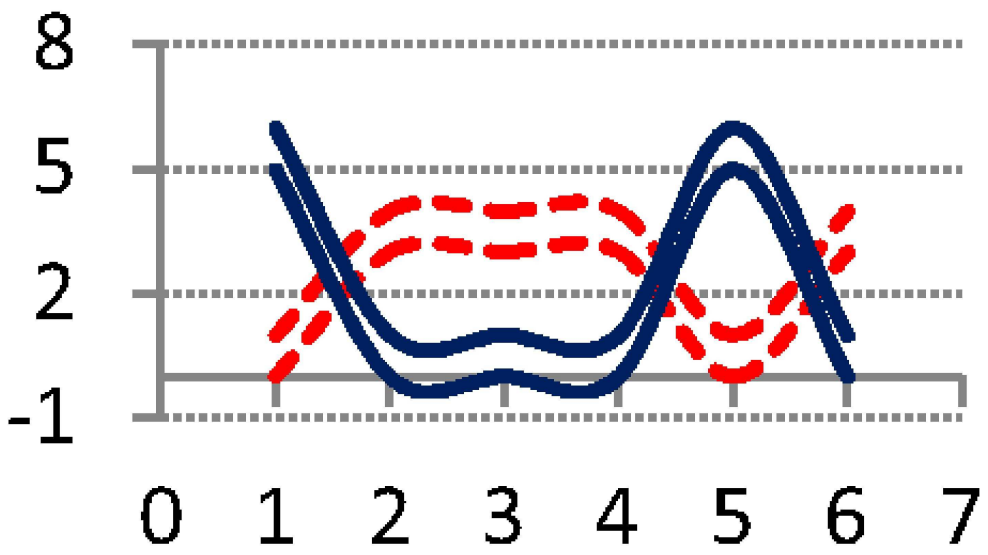


Figure 3.(a)
123x71mm (600 x 600 DPI)

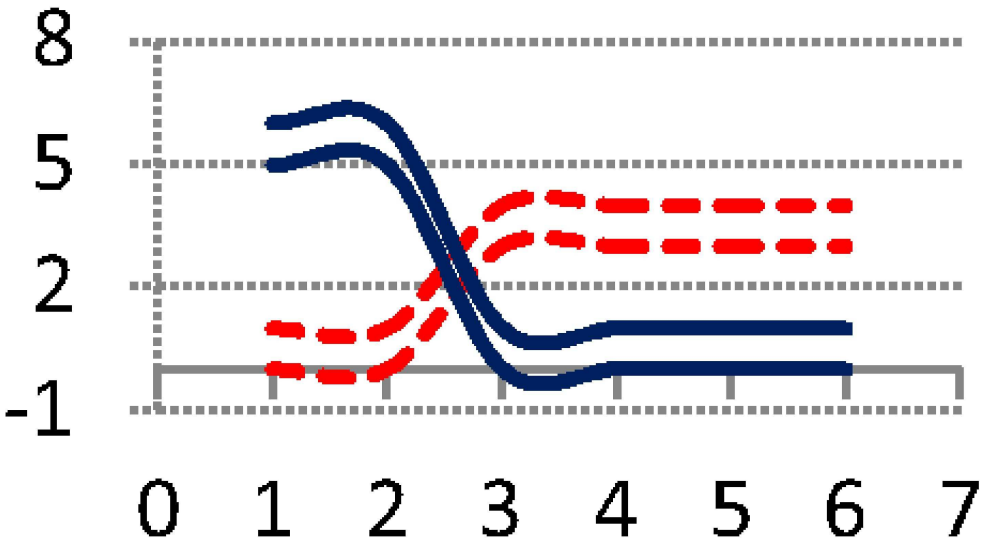


Figure 3.(b)
123x69mm (600 x 600 DPI)

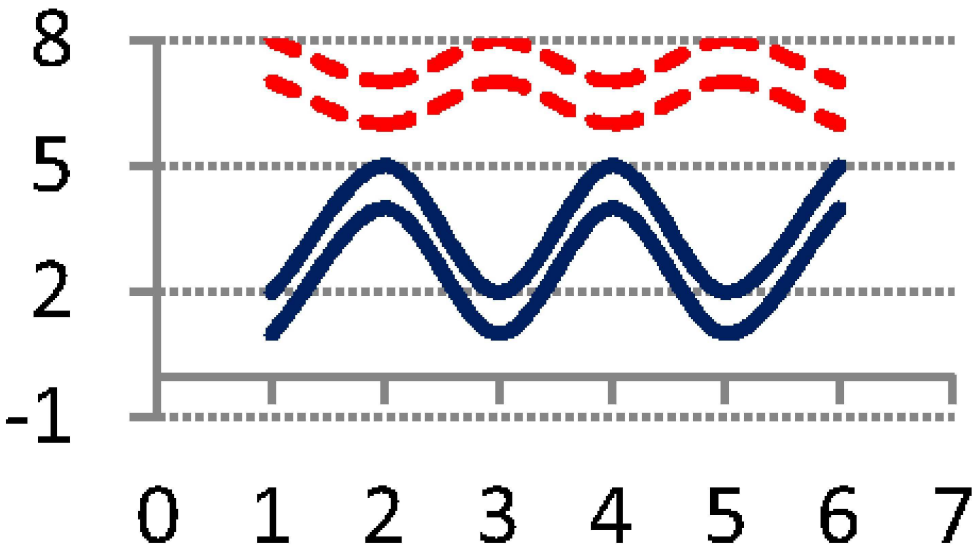


Figure 3.(c)
124x71mm (600 x 600 DPI)

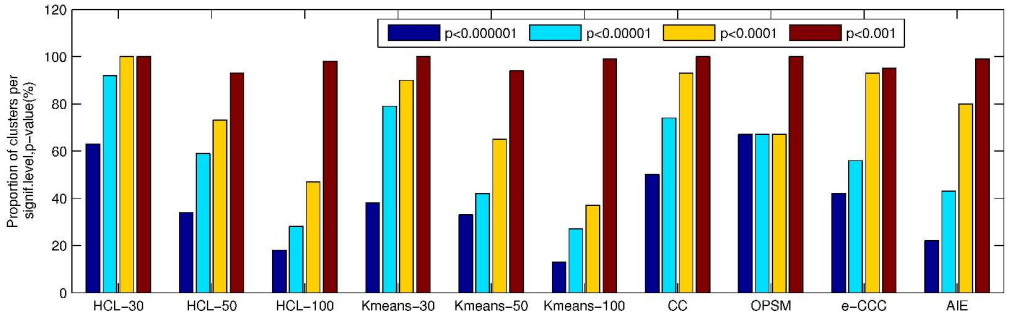


Figure 4.(a) Proportion of P-values in Alpha factor
205x63mm (600 x 600 DPI)

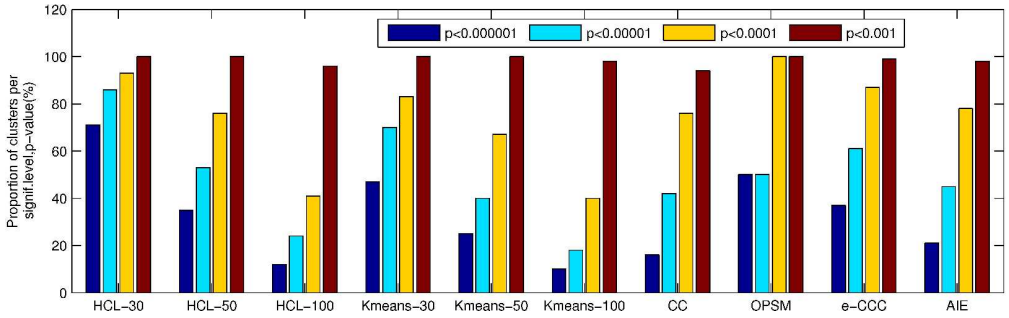


Figure 4.(b) Proportion of P-values in Cdc15
205x63mm (600 x 600 DPI)

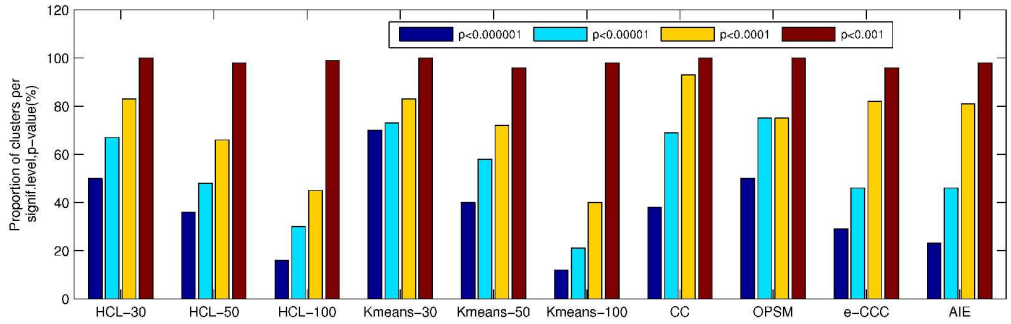


Figure 4.(c) Proportion of P-values in Cdc28
205x64mm (600 x 600 DPI)

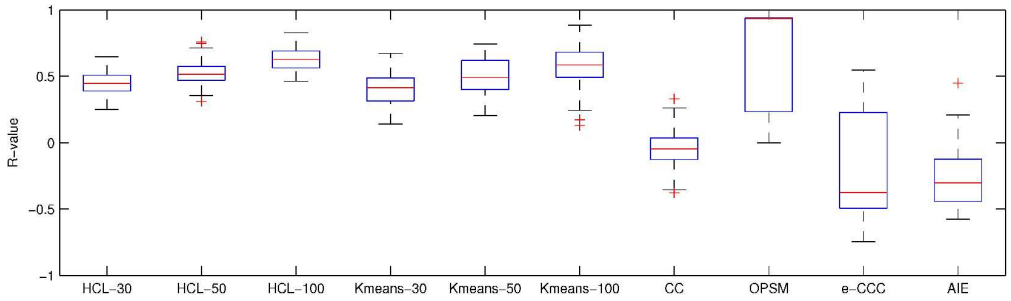


Figure 5.(a) R-values in Alpha factor
209x61mm (600 x 600 DPI)

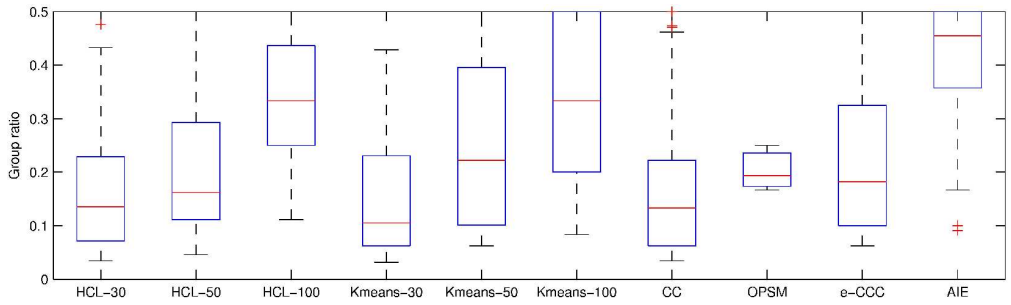


Figure 5.(b) Group ratio in Alpha factor
208x61mm (600 x 600 DPI)

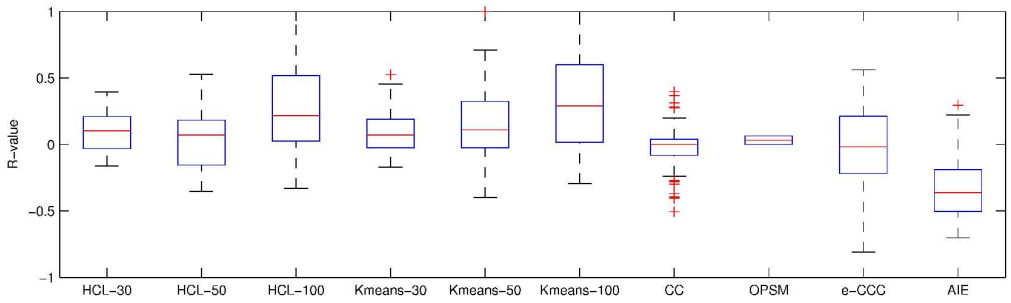


Figure 5.(c) R-values in Cdc15
209x61mm (600 x 600 DPI)

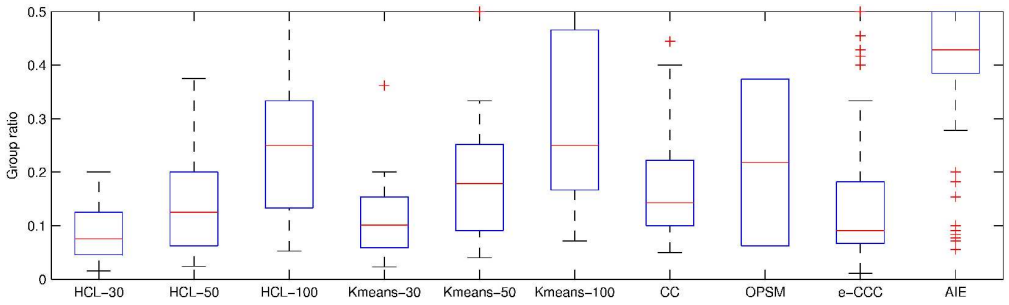


Figure 5.(d) Group ratio in Cdc15
208x61mm (600 x 600 DPI)

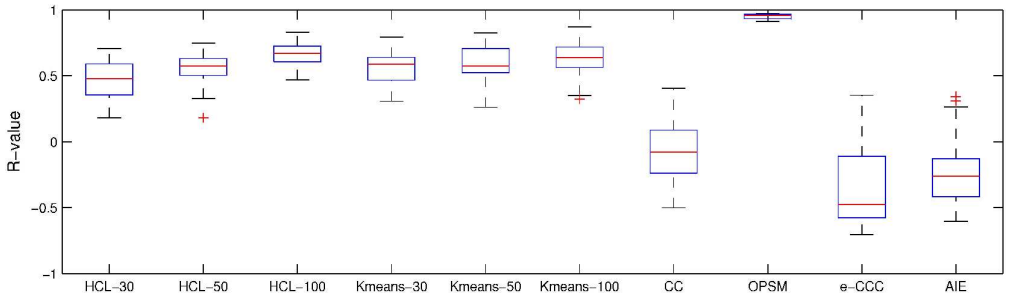


Figure 5.(e) R-values in Cdc28
210x61mm (600 x 600 DPI)

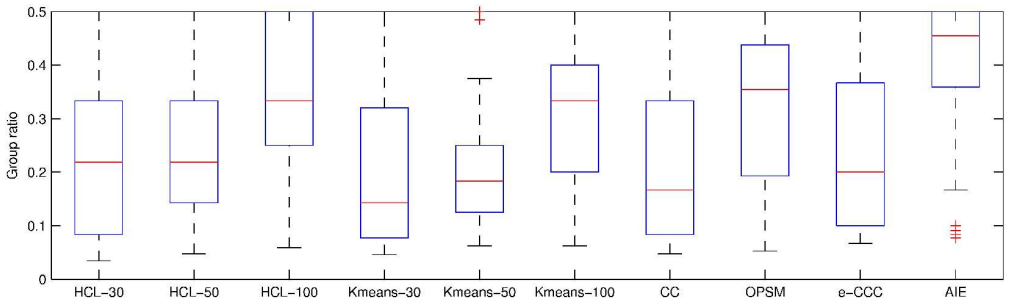


Figure 5.(f) Group ratio in Cdc28
208x61mm (600 x 600 DPI)

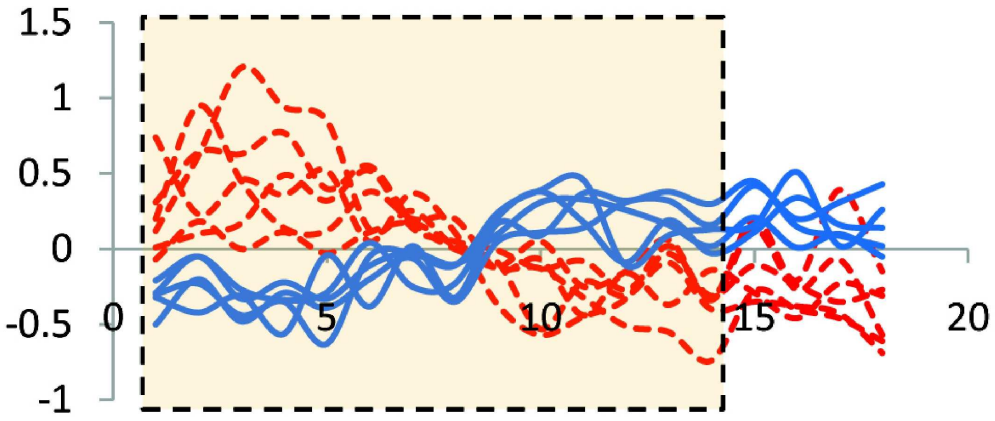


Figure 6.(a) From Alpha factor
181x74mm (600 x 600 DPI)

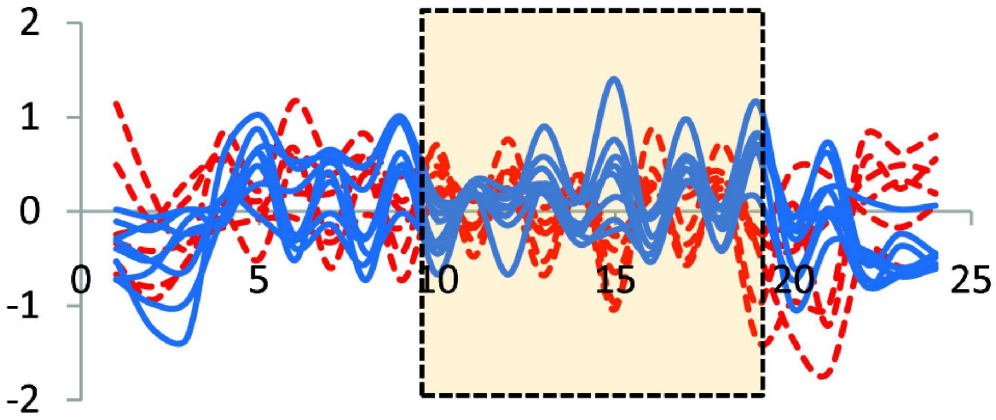


Figure 6.(b) From Cdc15
180x75mm (600 x 600 DPI)

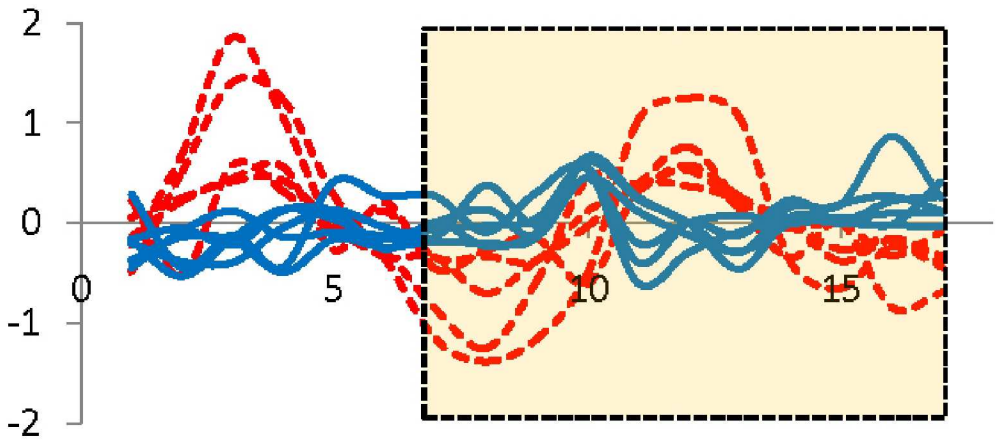


Figure 6.(c) From Cdc28
198x87mm (600 x 600 DPI)

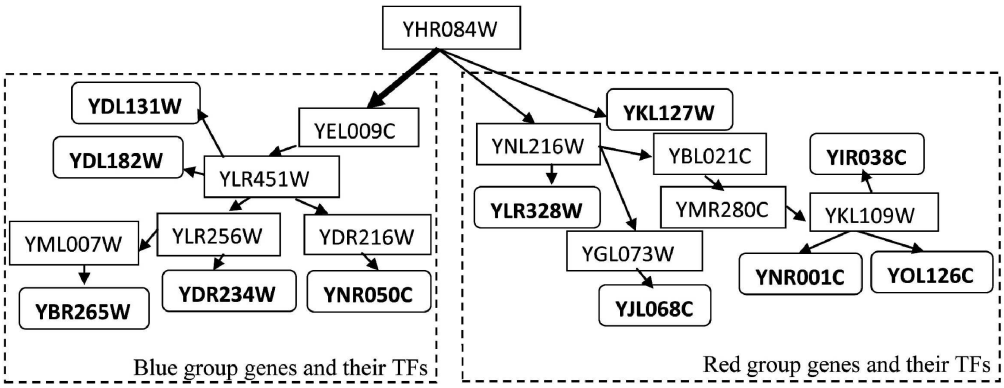


Figure 7. A sub-regulatory tree for AIE-Alpha-87.
196x75mm (600 x 600 DPI)

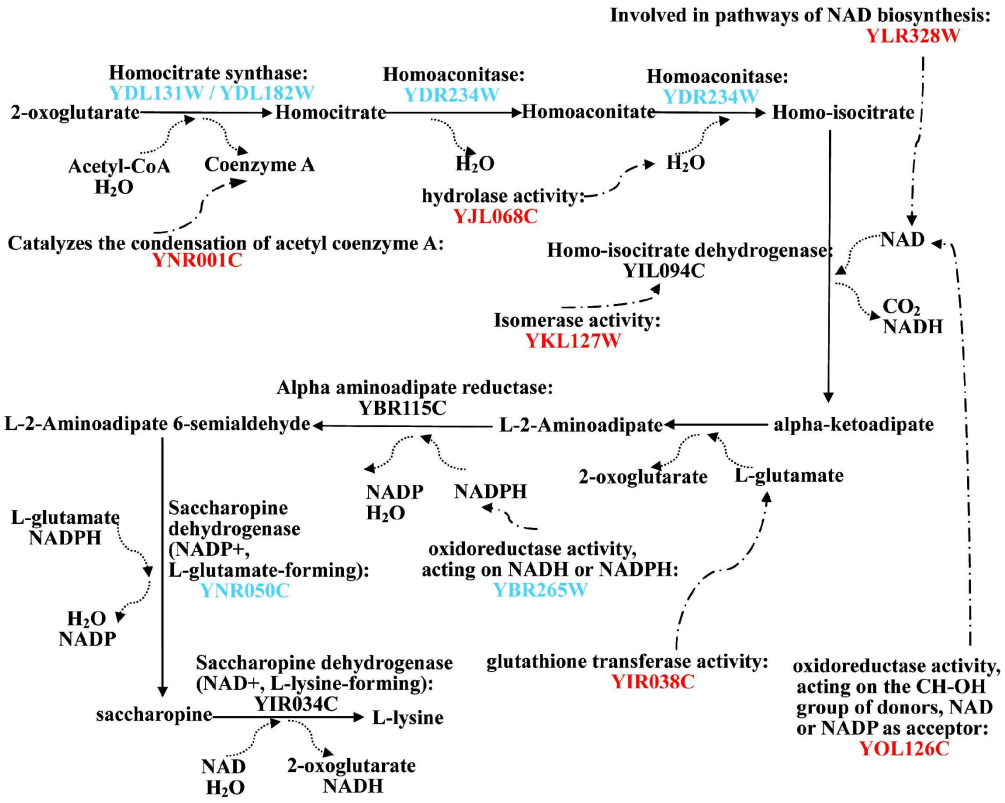


Figure 8. An expanded diagram for the pathway lysine biosynthesis after our functional annotation is used to derive new components (shown as dashed lines) based on the negative correlation of AIE-Alpha-87.
194x154mm (600 x 600 DPI)

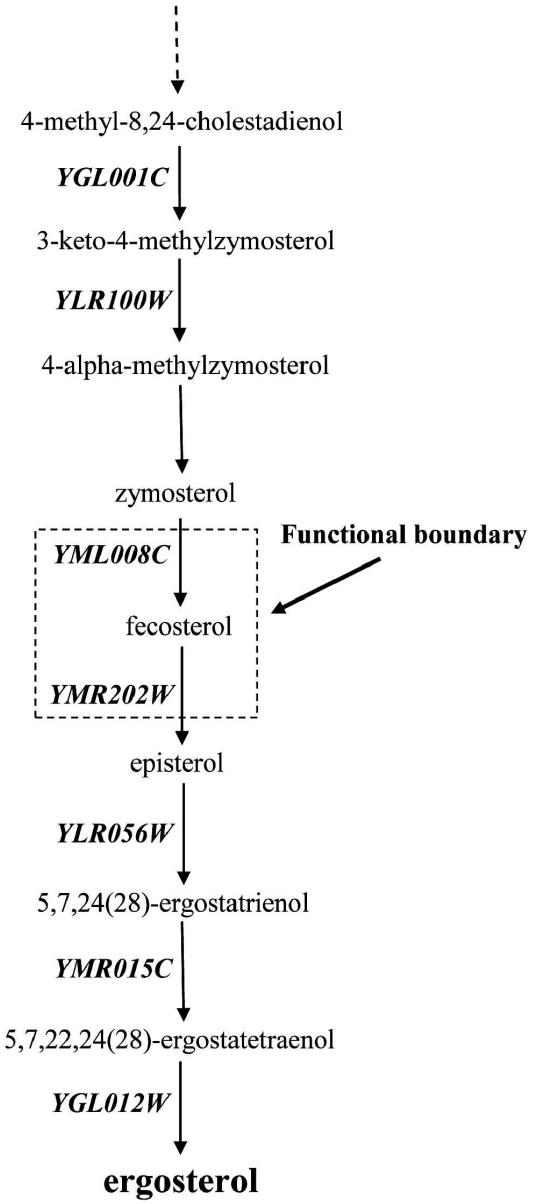


Figure 9. A partial diagram of the pathway ergosterol biosynthesis.
124x283mm (600 x 600 DPI)

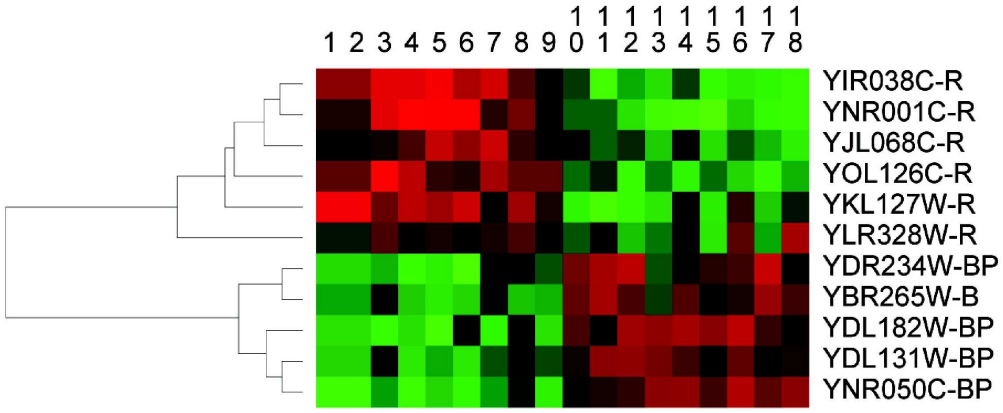


Figure 10.(a) AIE-Alpha-87 under Alpha factor
191x78mm (600 x 600 DPI)

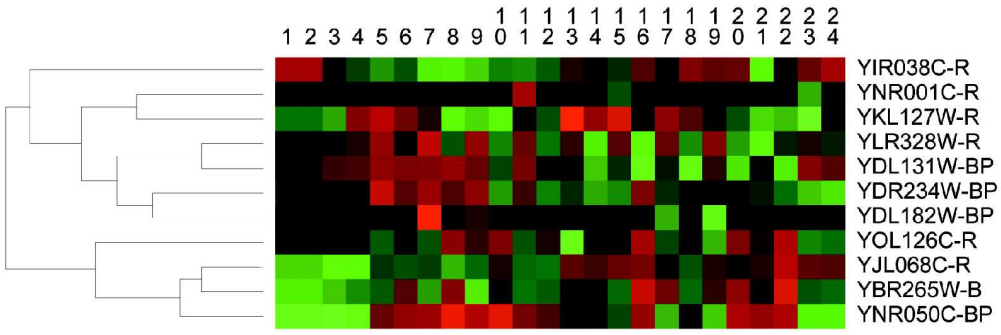


Figure 10.(b) AIE-Alpha-87 under Cdc15
201x66mm (600 x 600 DPI)

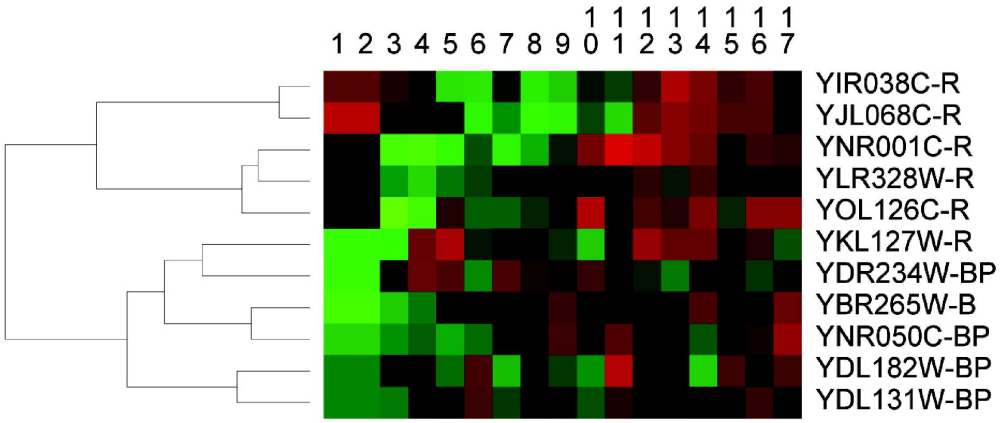


Figure 10.(c) AIE-Alpha-87 under Cdc28
191x80mm (600 x 600 DPI)

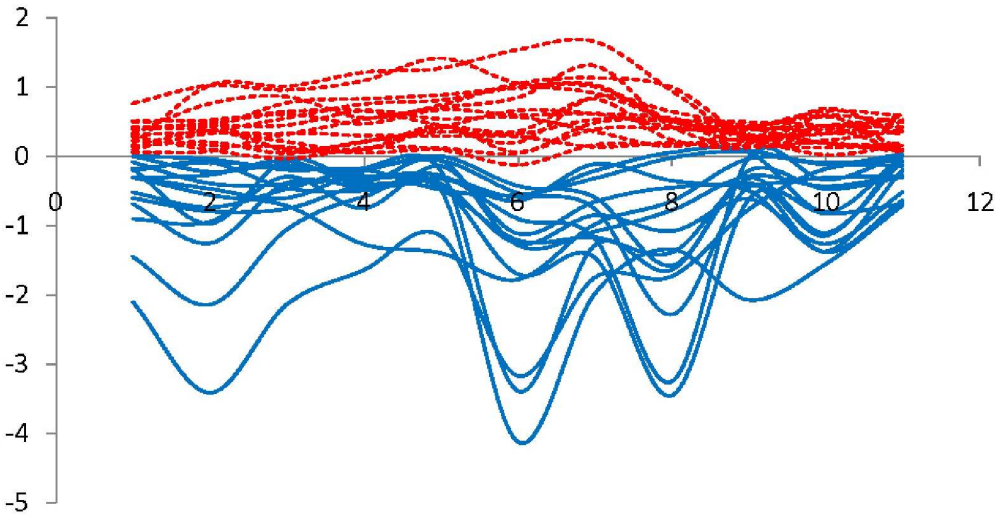


Figure 11.(a) AIE-BeforeStress-4411
201x103mm (600 x 600 DPI)

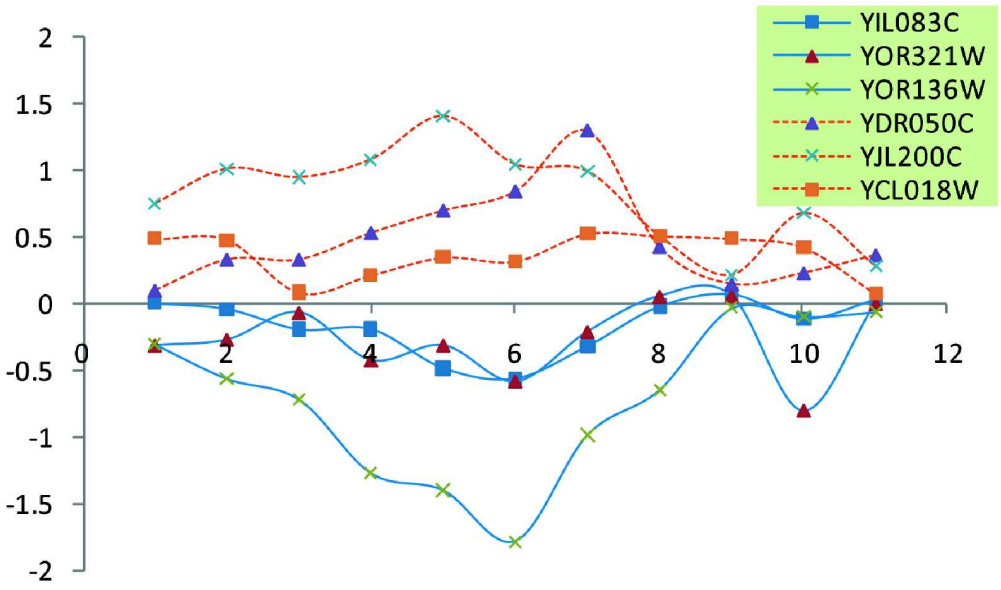


Figure 11.(b) Invariable component of
AIE-BeforeStress-4411
201x116mm (600 x 600 DPI)

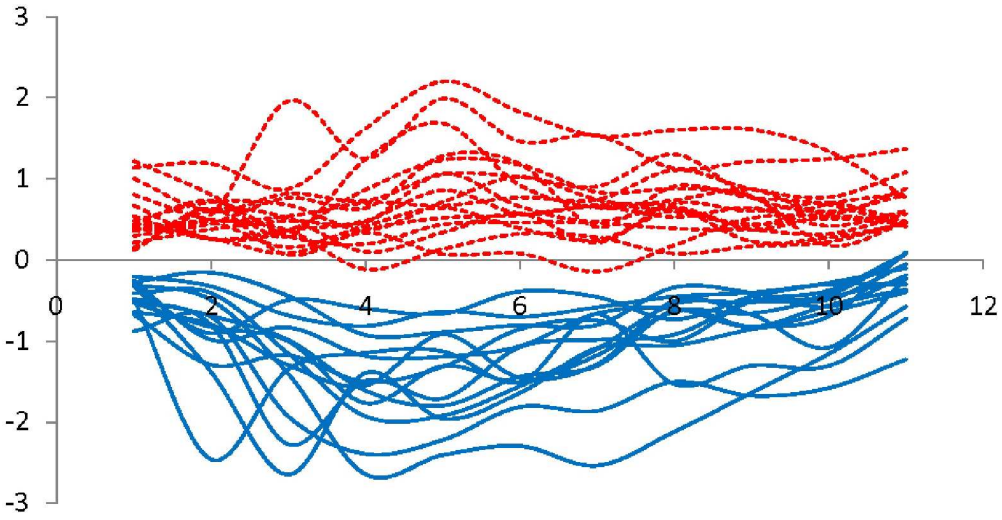


Figure 11.(c) AIE-AfterStress-3827
199x103mm (600 x 600 DPI)

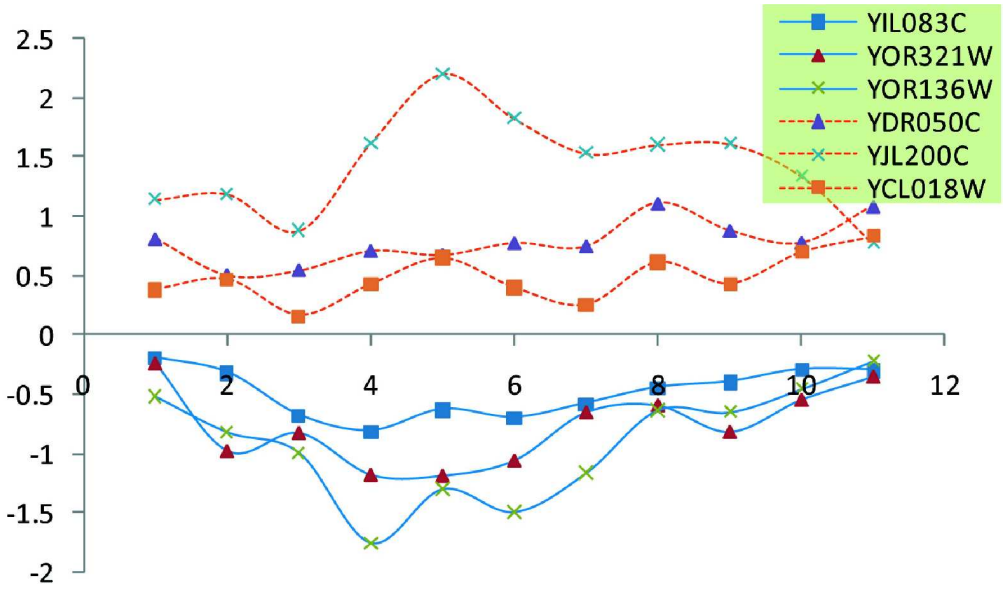


Figure 11.(d) Invariable component of
AIE-AfterStress-3827
200x116mm (600 x 600 DPI)