Gene expression

Towards patterns tree of gene coexpression in eukaryotic species

Haiyun Wang1,∗,†, Qi Wang2,∗,†, Xia Li1,3, Bairong Shen1,4,5, Min Ding1 and Ziyin Shen2

1School of Life Science and Technology, Tongji University, Shanghai 200092, 2Institute of Chinese Traditional Medicine and Western Medicine Integration, Huashan Hospital, Fudan University, Shanghai 200040, 3Department of Bioinformatics, Harbin Medical University, Harbin 150086, China, 4Center for Systems Biology, Soochow University, Suzhou 215006, China and 5Institute of Medical Technology, University of Tampere, Tampere 33014, Finland

Received on January 26, 2008; revised on April 8, 2008; accepted on April 9, 2008
Advance Access publication April 10, 2008
Associate Editor: Trey Ideker

ABSTRACT

Motivation: Cellular pathways behave coordinated regulation activity, and some reported works also have affirmed that genes in the same pathway have similar expression pattern. However, the complexity of biological systems regulation actually causes expression relationships between genes to display multiple patterns, such as linear, non-linear, local, global, linear with time-delayed, non-linear with time-delayed, monotonic and non-monotonic, which should be the explicit representation of cellular inner regulation mechanism in mRNA level. To investigate the relationship between different patterns, our work aims to systematically reveal gene-expression relationship patterns in cellular pathways and to check for the existence of dominating gene-expression pattern. By a large scale analysis of genes expression in three eukaryotic species, Saccharomyces cerevisiae, Caenorhabditis elegans and Human, we constructed gene coexpression patterns tree to systematically and hierarchically illustrate the different patterns and their interrelations.

Results: The results show that the linear is the dominating expression pattern in the same pathway. The time-shifted pattern is another important relationship pattern. Many genes from the different pathway also present coexpression patterns. The non-linear, non-monotonic and time-delayed relationship patterns reflect the remote interactions between the genes in cellular processes. Gene coexpression phenomena in the same pathways are diverse in different species. Genes in S.cerevisiae and C.elegans present strong coexpression relationships, especially in C.elegans, coexpression is more universal and stronger due to its special array of genes. However in Human, gene coexpression is not apparent and the human genome involves more complicated functional relationships. In conclusion, different patterns corresponding to different pathway also present coexpression patterns. The non-linear relationship may be time-shifted (Lee et al., 2000; Eisen et al., 1998; Niehrs and Pollet, 1999; Ucar et al., 2007; Wen et al., 1998; Zhang and Horvath, 2005). These analyses rest on the assumption that genes with similar expression profile function in the related biological processes (Eisen et al., 1998; Gerstein and Jansen, 2000; Marcotte et al., 1999; Niehrs and Pollet, 1999). Likewise, it also has been reported that the expression patterns for genes with similar functions are alike (Niehrs and Pollet, 1999; Waters, 2003).

A large number of coordinated genes are coordinated in biological processes and these coordinations show diverse patterns. For instance, some genes encoding regulators within the same category are bound by the same transcriptional regulators, and these genes may act in simultaneous expression (Lee et al., 2002). One gene may control or activate a downstream gene in a pathway and therefore their expression relationship may be time-shifted (Lee et al., 2002; Qian et al., 2001). Also, some genes can be dynamically coregulated, i.e. one gene may be coordinated with another along several specific time courses. In this case, their expression relationships could only indicate local similarity (Balasubramaniam et al., 2005; Filkov et al., 2002). Furthermore, an inhibitory relationship may exist between the genes, i.e. one gene’s down-expression could be induced by another gene’s up-regulation, and then we can expect their expression profiles to be inverted, respectively (Qian et al., 2001; Qin, 2006). In addition, expression patterns could also indicate non-monotonic relationships, i.e. early in the time course one gene’s expression level rises along with another’s, and, while one continues to rise, the other might level off in the middle state and fall in the late state. Thus, the complexity of biological processes is reflected in a great diversity of interrelationships between the genes, and these relationships may display various coexpression patterns.

1 INTRODUCTION

Biological systems involve many complex relationships which are very difficult to discover by the traditional biological methods. Thus, the high-throughput ‘-omics’ (proteomics, metabolomics, transcriptomics, genomics, etc.) technologies are gradually being adopted to systematic study of biological processes. Most of the current genome-wide expression data analyses focus on identifying correlated genes, extracting gene clusters and inferring the gene regulatory networks (Brown et al., 2000; Eisen et al., 1998; Niehrs and Pollet, 1999; Ucar et al., 2007; Wen et al., 1998; Zhang and Horvath, 2005). These analyses rest on the assumption that genes with similar expression profile function in the related biological processes (Eisen et al., 1998; Gerstein and Jansen, 2000; Marcotte et al., 1999; Niehrs and Pollet, 1999). Likewise, it also has been reported that the expression patterns for genes with similar functions are alike (Niehrs and Pollet, 1999; Waters, 2003).

A large number of coordinated genes are coordinated in biological processes and these coordinations show diverse patterns. For instance, some genes encoding regulators within the same category are bound by the same transcriptional regulators, and these genes may act in simultaneous expression (Lee et al., 2002). One gene may control or activate a downstream gene in a pathway and therefore their expression relationship may be time-shifted (Lee et al., 2002; Qian et al., 2001). Also, some genes can be dynamically coregulated, i.e. one gene may be coordinated with another along several specific time courses. In this case, their expression relationships could only indicate local similarity (Balasubramaniam et al., 2005; Filkov et al., 2002). Furthermore, an inhibitory relationship may exist between the genes, i.e. one gene’s down-expression could be induced by another gene’s up-regulation, and then we can expect their expression profiles to be inverted, respectively (Qian et al., 2001; Qin, 2006). In addition, expression patterns could also indicate non-monotonic relationships, i.e. early in the time course one gene’s expression level rises along with another’s, and, while one continues to rise, the other might level off in the middle state and fall in the late state. Thus, the complexity of biological processes is reflected in a great diversity of interrelationships between the genes, and these relationships may display various coexpression patterns.

© The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
such as linear, non-linear, local, global, linear with time-delayed, non-linear with time-delayed, monotonic and non monotonic. All these different coordination patterns should be the explicit representation of cellular inner regulation mechanism.

To investigate the coexpression patterns of genes in the cellular pathways and the relationship between these patterns, we carried out a large scale analysis of genes in three eukaryotic species: *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Human*, mapping species dependent genes onto the KEGG metabolic pathways and identifying multiple expression relationships of gene pairs using presented methods. Furthermore, we constructed gene coexpression patterns trees in different species to systematically and hierarchically organize and illustrate different patterns and their relationships. Through comprehensive patterns tree of genes coexpression, we can systematically understand relationships between genes and their expression activity in the cellular society.

2 ALGORITHM AND METHODS

We combined three different pattern scores, Pearson correlation coefficient (\(\gamma_p\)), Spearman rank correlation (\(\gamma_s\)) and mutual information (\(\gamma_m\)), to unravel linear, non-linear and non-monotonic relationship patterns. With the combination use of these simple and widely used correlation measures, all correlation relationships can be unraveled. Gene pairs with statistically significant scores using hypothesis test \(T\) were considered to have certain relationship patterns. If both \(\gamma_p\) and \(\gamma_s\) (or \(\gamma_p\) and \(\gamma_m\)) were significant for a given gene pair, their expression relationship was referred to as a linear pattern. If only \(\gamma_p\) or \(\gamma_s\) or \(\gamma_m\) were significant, the expression relationship was referred to as a non-linear pattern. Two randomization procedures were applied to detect the statistical significance of different relationship patterns.

2.1 Methods for unraveling gene coexpression patterns

The expression vectors of two genes, \(g_x\) and \(g_y\), are denoted by \(X\) and \(Y\), respectively. \(X = (x_1, x_2, ..., x_i, ..., x_n)\), \(Y = (y_1, y_2, ..., y_i, ..., y_n)\), \(x_i\) is the expression value of \(i\)-th experiment condition of \(g_x\), and \(y_i\) is the expression value of \(i\)-th experiment condition of \(g_y\).

If data are time course microarray experiments, time-delayed coexpression relationship should be considered. Suppose the expression profile of \(g_x\) is delayed by \(t\) time points against \(g_y\): \(X\) and \(Y\) can be denoted as \(X = (x_1, x_2, ..., x_{i-t}, ..., x_n)\) and \(Y = (y_1, y_2, ..., y_{i-t}, ..., y_{n-t})\), respectively. When the expression profile of \(g_x\) is delayed by \(t\) time points: \(X = (x_1, x_2, ..., x_i, ..., x_{n-t})\) and \(Y = (y_1, y_2, ..., y_i, ..., y_{n-t})\).

The basic definition of gene expression pattern score \(\gamma(x, y, \gamma_m)\) is described as following:

Pearson correlation coefficient:

\[
\gamma_p = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i\right)^2\right) \left(\sum_{i=1}^{n} y_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} y_i\right)^2\right)}}
\]

Where \(L\) is the dimension of \(X\) or \(Y\).

Spearman rank correlation:

\[
\gamma_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{L(L^2 - 1)}
\]

Here, we rank both \(X\) and \(Y\) from the highest to the lowest, then subtract the two sets of ranks to get the difference \(d\).

The mutual information of \(X\) and \(Y\) can be calculated by:

\[
\gamma_m = H(X) - H(X|Y)
\]

Where \(\gamma_m\) is the mutual information between \(X\) and \(Y\), \(H(X)\) is the entropy of \(X\), \(H(X|Y)\) is the conditional entropy of \(X\). When \(X\) and \(Y\) are discrete, they can be calculated by,

\[
H(X|Y) = -\sum_{i=1}^{m} p(x_i|y_j) \log p(x_i|y_j)
\]

\[
H(X) = -\sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i|y_j)p(x_i)p(y_j) \log p(x_i|y_j) \log p(x_i)
\]

Where \(p(.\)) is the probability density function which can be estimated by frequency. \(n\) and \(m\) are bins we use to discretize \(X\) and \(Y\).

For time-delayed gene pairs, \(\gamma\) corresponding to certain delayed time point is selected as the most possible time-delayed pattern.

2.2 Hypothesis testing procedure for score \((T)\)

We apply a test for pattern score with Monte Carlo techniques. For this hypothesis, we can test whether the calculated score \(\gamma(x, y, \gamma_m)\) for two genes is a random sample from the background distribution of scores while perturbing experiment conditions. The test procedure is as follows:

1. Create reference expression vectors of \(g_x\) and \(g_y\) under \(H_0\) by permuting experiment conditions of \(X\) and \(Y\), respectively.

2. Calculate pattern score \(\gamma_0\) of permuted \(X\) and \(Y\).

3. Repeat steps 1–2 for 500 times

4. Create histogram of \(\gamma_0\) (null distribution)

5. Calculate \(p(\gamma|H_0)\), if \(p < 0.05\), then reject \(H_0\)

Gene pairs with statistical significance are selected, and their scores reflect their co-expression patterns.

2.3 Randomization procedures

To identify the statistical significance of different patterns in real situation, two randomization procedures were applied. The first, which we name pseudo pathway simulation, randomly picks up some genes from the entire set of genes, but keeps the gene number equal to that in real pathway. The second procedure, which we name experiment condition permutation, only disturbs the experiment condition sequence of genes-expression vectors, but keeping genes mapped onto the real pathway.

The pseudo pathway simulates the relationships of genes from the different pathway and such relationship actually reflects a remote action between gene pairs. The experiment condition permutation procedure simulates absolute randomization in which gene pairs for the same pathway should have no coexpression patterns.

3 IMPLEMENTATION AND RESULTS

The microarray data analyzed here come from three eukaryotic species: *S.cerevisiae*(see), *C.elegans*(cel) and *Human*(hsa) (for details see Table 1).

The function pathways with at least two genes mapped in each were chosen for investigation after we mapped genes onto KEGG pathway. Gene pairs, each of which being from some one pathway, were calculated with different methods to identify their relationships.
The patterns distribution in different situations

<table>
<thead>
<tr>
<th>Species</th>
<th>Situation</th>
<th>Linear</th>
<th>Non-linear</th>
<th>Non-monotonic</th>
<th>No relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>see</td>
<td>Real situation</td>
<td>11733 (56.72%)</td>
<td>1260 (6.09%)</td>
<td>1233 (5.96%)</td>
<td>6461 (31.23%)</td>
</tr>
<tr>
<td></td>
<td>Pseudo pathway simulation</td>
<td>8746 (42.28%)</td>
<td>1149 (5.55%)</td>
<td>1094 (5.29%)</td>
<td>9698 (46.88%)</td>
</tr>
<tr>
<td>cel</td>
<td>Real situation</td>
<td>15370 (63.51%)</td>
<td>759 (3.14%)</td>
<td>1539 (6.36%)</td>
<td>6532 (26.99%)</td>
</tr>
<tr>
<td></td>
<td>Pseudo pathway simulation</td>
<td>14528 (60.03%)</td>
<td>889 (3.67%)</td>
<td>1702 (7.03%)</td>
<td>7081 (29.26%)</td>
</tr>
<tr>
<td>hsa</td>
<td>Real situation</td>
<td>12659 (56.72%)</td>
<td>4998 (8.73%)</td>
<td>3974 (6.94%)</td>
<td>6461 (31.23%)</td>
</tr>
<tr>
<td></td>
<td>Pseudo pathway simulation</td>
<td>8746 (42.28%)</td>
<td>1149 (5.55%)</td>
<td>1094 (5.29%)</td>
<td>9698 (46.88%)</td>
</tr>
</tbody>
</table>

*Denotes the number of pathways with at least two genes mapped.

For each species, number (percent) of gene pairs with linear, non-linear, non-monotonic and no relationship pattern is shown.

Table 2. Dataset information

<table>
<thead>
<tr>
<th>Species</th>
<th>Dataset</th>
<th>Genes</th>
<th>Samples</th>
<th>Data type</th>
<th>Number of pathways (*)</th>
<th>Gene pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>see</td>
<td>GSE3431</td>
<td>6551</td>
<td>36</td>
<td>Metabolic cycle time course</td>
<td>105 (78)</td>
<td>20687</td>
</tr>
<tr>
<td>cel</td>
<td>GSE9967</td>
<td>18649</td>
<td>20</td>
<td>3 Wildtype and 9 mutants</td>
<td>118 (65)</td>
<td>24200</td>
</tr>
<tr>
<td>hsa</td>
<td>NC160</td>
<td>6830</td>
<td>64</td>
<td>60 Cancer cell lines</td>
<td>197 (184)</td>
<td>37264</td>
</tr>
</tbody>
</table>

3.1 Discovering different relationship patterns

Gene pairs with statistically significant $t$ value under hypothesis testing procedure T were selected. Table 2 shows the patterns distribution in real situation and pseudo pathway simulation. The patterns distribution in the same pathway appeared to be different in three species. *C.elegans* contains 73.01% (63.51% + 3.14% + 6.36%) gene pairs that present coexpression patterns, while *S.cerevisiae* and *Human*, respectively, contains 68.77% (56.72% + 6.09% + 5.96%) and 37.77% (22.11% + 8.73% + 6.94%). Gene coexpression does widely occur in *C.elegans* and *S.cerevisiae* while in *Human* apparently it does not. Gene coexpression also presents in the genes from the different pathway. Moreover, linear relationship is the dominating expression pattern both for the same pathway and the different pathway in three species.

Figure 1A displays the distributions of pattern scores in *S.cerevisiae*. A linear relationship between genes is seen (left panel in Fig. 1A) and the proportion of gene pairs with a score >0.5 or <-0.5 in real situations is significantly more than that in randomization. This implies that the gene pairs in the same pathway are inclined to have linear relationship pattern against two randomized situations. Also, the majority of score in the real situation is positively biased, and another obvious peak with the center of 0.7 appears, suggesting that the majority of gene pairs in the same pathway prefer to regulate positively rather than negatively. Moreover, although the shapes of both two randomized situations are close to normal distribution, the peak for experiment condition permutation is obviously sharper and narrower than that of pseudo pathway simulation. The experiment condition permutation procedure simulates genes pairs of no co-expression relationship, so the shape of its curve is characterized by similar normal distribution with zero median and small variance. However, the pseudo pathway actually simulates a remote action between two genes from the different pathway. As the left panel in Figure 1A shows, the shape of pattern score in the pseudo pathways situation is closer to that for the real situation than that in experiment condition permutation. Gene pairs with a relatively high absolute value of Pearson correlation in the pseudo pathways situation are also significantly more than that of experiment condition permutation, which means that gene pairs with the remote action also have linear coexpression relationship to a certain extent, but such coexpression relationships are not as universal as gene pairs from the same pathways.

Figure 1B displays the distributions of pattern scores in *C.elegans*. Left panel in Figure 1B shows the majority of linear score in the real situation is positively biased, which also indicates *C.elegans* genes tend to regulate positively. Moreover, a larger proportion of gene pairs in *C.elegans* present linear patterns both in real and pseudo pathways, and the distribution shape in real situation differs greatly from that of pseudo pathway in appearing an obvious peak near 1.0, suggesting that in *C.elegans* gene linear coexpression pattern is ubiquitous both in the same pathway and different pathway and the highly correlated gene pairs tend to act in the same pathways.

Figure 1C displays the distributions of pattern scores in *Human*. Surprisingly, just as left panel in Figure 1C shows, there is no obvious difference between real situation and pseudo pathway, and both situations are close to normal distribution with zero median, which indicates in *Human* gene coexpression in the same function is not apparent.

We further investigated non-linear relationship between genes by observing the distribution of Spearman correlation. As the middle panel in Figure 1A, B and C shows, the distribution of Spearman correlation in real situation is almost identical to that in pseudo pathway simulation, indicating that...
the non-linear relationship is not a specific pattern for genes in the same pathway, whereas it significantly exists for genes from the different pathways against absolute randomization distribution. Actually, the non-linear relationship should be the real reflection of remote action between genes from both the different pathways and the same pathways.

To address non-monotonic relationship between genes, we also determined the distribution of mutual information in real situation, pseudo pathway simulation, and experiment condition permutation (right panel in Fig. 1A, B and C). The distribution shape for the experiment condition permutation is sharp and peaks closer to the zero value, while the real situation distribution, being similar to the distribution of the pseudo pathway simulation, shows a higher peak value of the mutual information. Altogether, this indicates that the non-monotonic relationship between the genes in the same pathway or from the different pathways is more universal than that in an absolutely randomized situation.

3.2 Discovering time-delayed relationship patterns
For time course experiment, gene pairs may present time-shifted relationships besides simultaneous coexpression. In S.cerevisiae, 31.23% gene pairs did not show significant relationships. Hence these gene pairs were further analyzed to detect possible time-delayed relationship pattern.

Briefly, the time-delayed relationship analysis carried out as follows. Suppose the delayed time between two gene expressions is \( t \). Parameter \( t \) was changed from 1 to 12, since 36 time points were measured in the microarray experiment and many of the genes exhibited robust, highly periodic cycles with 12 time intervals per cycle (Tu et al., 2005). For each pair of genes, the highest scores corresponding to certain delayed time point were selected to do hypothesis test \( T \), and the gene pairs with statistically significant scores and delayed pattern were identified.

A total of 3354 gene pairs (16.21%) showed significant time-delayed expression relationship and among these 2559 pairs
showed linear patterns, 522 non-linear patterns and 273 showed non-monotonic patterns.

The distributions of different time-delayed pattern scores in real situation are very similar to ones in pseudo pathway simulation (Fig. 2), meaning that time-delayed patterns are equally present both for the genes from the same pathway and the genes from the different pathways. Differing from the plots in Figure 1, the distributions of the time-delayed Pearson correlation (Fig. 2A) and Spearman correlation (Fig. 2B) show two distinct peaks centering around $\pm 0.5$, because more than half of the gene pairs that did not show simultaneous patterns were found to have time-delayed relationship. By adjusting the delayed time points, many of the pairs showed the relatively high time-delayed pattern scores (around $\pm 0.5$).

### 3.3 Gene coexpression patterns trees

According to our analysis of gene coexpression patterns, we constructed genes coexpression patterns tree for each species (Fig. 3), to systematically and hierarchically reveal gene pairs’ relationships.

In *C.elegans* and *Human*, total gene pairs can be classified into two groups: coexpressed and no relationship (Fig. 3B and C). Coexpressed pattern can be further sub-grouped to monotonic and non-monotonic patterns, and monotonic pattern could be divided into the ‘linear’ and ‘non-linear’ relationship nodes. In *C.elegans*, majority of genes in the same pathways are coexpressed, while in *Human*, only a small proportion of genes are coexpressed.

All gene coexpression relationships in *S.cerevisiae* can be classified into three groups (Fig. 3A): simultaneous expression relationship, time-delayed expression relationship and no relationship. The root node indicates that the simultaneous relationship patterns (blue) occupy 68.77% of all gene pairs and the proportion of time-delayed patterns (red) is 16.21%. Simultaneous pattern and time-delayed pattern can be further sub-grouped to monotonic and non-monotonic patterns. 91.33% of all simultaneous pairs and 91.86% of all time-delayed pairs have monotonic pattern, indicating that monotonic pattern is the main expression relationship for gene pairs in the same pathway. Furthermore, monotonic pattern could be divided into the ‘linear’ and ‘non-linear’ relationship nodes. 90.30% of all monotonic pattern pairs in the ‘simultaneous’ node and 83.06% of all monotonic pattern pairs in the ‘time-delayed’ node showed linear relationship, respectively. Thus the relationship of linear with simultaneous pattern could be the dominating pattern for the gene pairs in the same pathways and the linear with time-delayed relationship is another important regulation event for cellular processes.

We further analyzed the correlation degree of the gene pairs with simultaneous linear pattern in *S.cerevisiae*.

According to the definition of statistics, the $\gamma_p$ scores were divided into three levels: high correlation ($\gamma_p \geq 0.7$), moderate correlation ($0.7 > \gamma_p \geq 0.4$), and low correlation ($0.4 > \gamma_p \geq 0.3$). Among the pairs 31.44% showed high
correlation, more than half correlated moderately and only 14.21% showed low correlation. In order to identify local correlation relationships, we selected two genes with low correlation, calculated the maximum Pearson correlation of continuous sub-time points for given time point length from 6 to 36, then observed the relationship between the length of time point and the absolute value of Pearson correlation. The result shows that the absolute value of Pearson correlation is descending with the time, but there is a distinct inflexion at the time point 24 (indicated by the arrow in Supplementary Fig. S1). At this point, Pearson correlation of gene pairs exceeds 0.6. Therefore, local similarity relationship is proved to exist. Accordingly, we extended the coexpression patterns tree (Fig. 3) with two kinds of leaf nodes, ‘local’ and ‘global’. When two genes are globally correlated, they are expected to have a high global gene pattern score. Otherwise, when the genes are locally correlated, their global pattern score will be relatively low while their local pattern score may be high.

4 DISCUSSION

The present analysis shows that in S.cerevisiae and C.elegans, 84.98% (68.77% + 16.21%) and 73.01% gene pairs in the same pathway are coexpressed, supporting the idea that genes from the same cellular pathway will display similar expression pattern. Also, just as left panel in Figure 1A and B show, the proportion of positive scores is obviously higher than that of negative scores, indicating that genes in the same pathway are disposed to be co-regulated positively rather than negatively. This conclusion is consistent with other studies. When using the Munich Information Center for Protein Sequences (MIPS) classifications, Cohen et al. (2000) found that S.cerevisiae genes with similar functions tended to occur in adjacent positions along the chromosomes, while adjacent genes in all orientations tended to be coexpressed. Moreover, according to Fukuoka et al. (2004), in yeast only the highly positively correlated pairs shared the same category and none of the pairs shared the same category in the zero-correlated and negatively correlated pairs. However in Human, more variety was obtained and pairs having the same category were found even in the zero-correlated and negatively correlated pairs. Our results also support the results of Cohen et al. (2000) concerning human. Just as left panel in Figure 1C shows, there is no obvious difference between real situation and pseudo pathway, and no positive bias for shape in real situation, which indicates the coexpression for genes in the same pathway is not apparent in Human and the human genome is more complicated.

In C.elegans, a large proportion of gene pairs present linear patterns both in real and pseudo pathways, and the distribution shape in real situation differs greatly from that of pseudo pathway in appearing an obvious peak near 1.0. High coexpression phenomena in C.elegans may be due to its special array of genes. The nematode worm C.elegans and its relatives are unique among animals in having operons. About 15% of all genes are in the operons (Blumenthal et al., 2002). Moreover, the researches of Lercher et al. (2003) suggest that the genome organization of C.elegans differs from the genomes of other eukaryotes not only by the existence of operons, but also by the relative role played by recent gene duplications. It is known that the genome of C.elegans contains many pairs of duplicated genes (Semple and Wolfe, 1999). Obvious peak near 1.0 may suggest that the highly correlated gene pairs, such as genes in the same operons or duplicated genes, tend to act in the same pathways.

Linear relationship is dominant in three species compared with other coexpression relationship and gene pairs within the same pathways have higher positive Pearson correlation than those from the different pathways in S.cerevisiae and C.elegans (left panel in Fig. 1A and B), thus Pearson correlation can unravel most of gene coexpression relationship in the same pathway. This demonstrates that the Pearson correlation coefficient (PCC) or its modification can be used as effective and the main measurement method for defining gene–gene relationships (Bergmann et al., 2004; Carter et al., 2004; Eisen et al., 1998; Stuart et al., 2003; Yeung et al., 2004).

Researchers have attempted to predict protein function by expression clustering. This is based on ‘guilt by association’ (Altman and Raychaudhuri, 2001), the premise that genes with similar expression profiles have similar functions or function in the same pathway. However, our results show that this hypothesis is not as powerful as expected, since gene coexpression relationships are not only apparent for the genes within the same pathway, but also exist for genes in the different pathways which was mimicked by the pseudo pathway simulation. For such a prediction about the relationship between expression and functional similarity, it is necessary to have a rigid limitation to some conditions. For example in S.cerevisiae, since the density of scores in the real situation, being greater than 0.5, are higher than those in the pseudo pathway simulation, we suggest that when PCC is used to assess the gene expression similarity, 0.5 may be a reasonable threshold for the data. Moreover, in Human this hypothesis is not appropriate because genes in the same pathway do not present similar expression patterns.

A large proportion of gene pairs in the pseudo pathway simulation were coexpressed (Table 2). Therefore, the harmonious behavior exists widely also between genes from the different pathways, and the cellular society is surely a complex biological system that is composed of a great many genes with coordinated regulation activity. There are no apparent differences of the non-dominating patterns between real situation and pseudo pathway simulation, therefore these non-dominating patterns may be the explicit representation of remote interactions between the genes in cellular processes.

Considering the metabolic cycle time course may be the external parameters to generate correlation, we applied another dataset of S.cerevisiae (GSE5938) including 70 microarrays under wildtype and genetic perturbations conditions. Patterns distributions based on two datasets are very similar (Supplementary Table S1), and therefore patterns tree that was constructed by our methods is relatively robust.

The gene regulatory networks are commonly studied by expression clustering from microarray data. According to our analysis of S.cerevisiae, coexpression of 16.21% of gene pairs in the same pathway is time-delayed while 18.16% of gene pairs in the different pathways show time-delayed coexpression. Thus, time-delayed coexpression pattern is also an important relationship. Consistently, the time-delayed gene
regulation in organisms is indeed a common phenomenon (Yeang, 2003). Hence, to the reconstruct gene regulatory networks based on the expression profiles, it is necessary to infer a time-delayed phenomenon in gene regulations by developing time-delayed relationship scores.

We checked the annotation of genes with high correlation. For instance in *S cerevisiae*, (1) gene pair of YJR063W and YNL248C is typically linear with simultaneous pattern. Both of genes are components of RNA polymerase transcription complex (Supplementary Fig. S2A). (2) YBL039C encodes the major CTP synthase isozyme and is the raw materials for RNA synthesis which is catalyzed by the YJR063W gene product (Supplementary Fig. S2B). (3) YDL024C and YDR481C were found to have a non-monotonic with simultaneous pattern. As Supplementary Figure S2C shows, when the expression level of YDR481C is relatively low (<3), the expression level of YDL024C gradually increases. While when mRNA expression level of YDR481C is relatively abundant (>3), the expression level of YDL024C gradually decreases. The products of YDL024C and YDR481C are involved in the gamma-hexachlorocyclohexane degradation pathway, both being phosphatases that act in the same reaction transforming 4-nitrophenylphosphate into 4-Nitrophenol. The co-expression pattern of these two genes reflects their intrinsic self-offset coordinate behavior serving for the same biological reaction. YDL024C encodes a protein of unknown function. Our analysis predicts that the YDL024C product may be an adjuvant enzyme produced increasingly to make up for the YDR481C phosphatase insufficiency and produced decreasingly until the supply of YDR481C enzyme is sufficient for this reaction. (4) YPR160W and YJL092W have linear with time-delayed pattern. The expression of YPR160W is three times delayed compared with that of YJL092W. In the upper part of Supplementary Figure S2D, there is no expression relationship between these genes, but the lower part shows a linear with three times delayed pattern unraveled by our method. YPR160W and YJL092W are both involved in the starch and sucrose metabolism pathway, and they are required for maintaining the normal level of α-D-Glucose. In conclusion, different patterns reflect different coordinating behaviors needed in maintaining the normal functions in the whole system. Different kinds of coordination patterns should be the representation of the cellular inner regulation mechanism. The gene coexpression patterns tree gives a comprehensive insight and understanding of the gene expression activity in cellular society.

ACKNOWLEDGEMENTS

The authors are most grateful to Huovila AP for the modification and suggestion in this article. They also thank YangY, Cheng YQ, Yao Y and Zhao X for contributions to this work.

Conflict of Interest: none declared.

REFERENCES


