Gene expression

MARD: a new method to detect differential gene expression in treatment-control time courses

Chao Cheng¹,²,*, Xiaotu Ma¹, Xiting Yan¹, Fengzhu Sun¹ and Lei M. Li¹,²,∗
¹Molecular and Computational Biology Program, Department of Biological Sciences, Computational Biology, University of Southern California Los Angeles, CA, USA and
²Department of Mathematics, University of Southern California Los Angeles, CA, USA

Received on April 25, 2006; revised on July 27, 2006; accepted on August 17, 2006
Advance Access publication August 23, 2006
Associate Editor: John Quackenbush

ABSTRACT

Motivation: Characterizing the dynamic regulation of gene expression by time course experiments is becoming more and more important. A common problem is to identify differentially expressed genes between the treatment and control time course. It is often difficult to compare expression patterns of a gene between two time courses for the following reasons: (1) the number of sampling time points may be different or hard to be aligned between the treatment and the control time courses; (2) estimation of the function that describes the expression of a gene in a time course is difficult and error-prone due to the limited number of time points. We propose a novel method to identify the differentially expressed genes between two time courses, which avoids direct comparison of gene expression patterns between the two time courses.

Results: Instead of attempting to ‘align’ and compare the two time courses directly, we first convert the treatment and control time courses into neighborhood systems that reflect the underlying relationships between genes. We then identify the differentially expressed genes by comparing the two gene relationship networks. To verify our method, we apply it to two treatment-control time course datasets. The results are consistent with the previous results and also give some new biologically meaningful findings.

Availability: The algorithm in this paper is coded in C++ and is available from http://leili-lab.cmb.usc.edu/yeastaging/projects/MARD/
Contact: lilei@usc.edu; chaochen@usc.edu
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Microarray techniques have been widely applied to identify genes that have different expression under various biological conditions. In many cases, we regard one condition as treatment and the other condition as control, which leads to the definition of treatment-control experiment design. Differential expression between the treatment and control condition can be investigated from either a static or temporal viewpoint. In a static experiment design, snapshots of gene expression levels are taken without considering the temporal effect; While in a temporal experiment design the gene expression across several time points (Wichert et al., 2004). For example, clustering analysis is often performed on a time course data to identify gene clusters with interesting expression patterns (Eisen et al., 1998; Luan and Li, 2003). On the other hand, several approaches have been proposed to compare time courses and identify differentially expressed genes between them. If the sampling time points can be ‘aligned’ between the treatment and control time courses, we can identify differentially expressed genes by direct comparison of the gene expression patterns under the two conditions. Available methods include the fold-change analysis (Yoshimoto et al., 2002), order-restricted statistics (Peddada et al., 2003), the ANOVA (Park et al., 2003), and one-sample multivariate empirical Bayes statistic (Tai and Spud, 2006).

However, two difficulties exist in the analysis of microarray data from a temporal design. First, the sampling time points are generally different from one study to another (Bar-Joseph, 2004). As a consequence, it is hard to integrate data from different studies. Second, a treatment may alter the ‘life-clock’ pace of an organism. For example, it has been reported that the knockout of gene sch9 extends yeast life-span by 3 folds (Fabrizio et al., 2001). In this case it is difficult to align the time-scales of individuals under treatment and control conditions. When the sampling time points cannot be ‘aligned’ between treatment and control, various interpolation techniques are often used as a preprocessing step before the direct comparison. For example, Bar-Joseph et al. (2003) proposed to represent gene expression patterns in treatment and control by B-spline curves and then compute a global difference measure between these two curves. Recently, Storey et al. (2005) proposed to represent gene expression trajectories by a natural cubic spline and then use goodness-of-fit test for differentiation detection. However, in either situation, the time points of the two time courses must be ‘aligned’.

In this paper, we propose a novel method to identify differentially expressed genes between control and treatment time courses which do not require ‘aligned’ time points in the two time courses. The method is proposed for the following considerations. (1) The relationships between genes can be estimated from microarray time course data. Functionally, associated genes tend to have similar expression patterns. So we can construct a gene relationship network from a microarray dataset, where each node is a gene and each
edge links two genes with similar expression patterns. Two gene relationship networks can be constructed from control and treatment time courses that may be different from each other. (2) Due to the robustness of cell system (Li et al., 2004; Alon et al., 1999), we may expect the gene relationship network to be also robust. Namely, we may expect that the majority of gene relationships are only marginally affected by a nonlethal treatment. Otherwise, dramatic changes in the whole relationship network may cause lethal effect. (3) If a gene is substantially affected by a treatment, we would expect a dramatic change of the gene between the two relationship networks constructed from control and treatment time courses. So we estimate the effect of the treatment on a gene indirectly by investigating the gene’s neighborhood change between the two relationship networks constructed from the treatment and control time courses. For example, if the neighbors of a gene in the two networks change dramatically, we regard this gene as substantially affected by the treatment. Based on these three considerations, we design a statistic called MARD (Mean Absolute Rank Difference) to measure the effect of a treatment on a gene. Since we compare the two constructed relationship networks instead of directly comparing the expression profiles of gene i under the control and treatment condition, respectively.

We apply our method to two treatment-control time course microarray datasets. One dataset is from an experiment with matched sampling time points while the other is from two data sources where the sampling time points cannot be well matched. In both datasets, we show that our approach identifies a set of genes that are well consistent with previous literatures. Furthermore, the approach also identifies a set of new genes in the corresponding biological contexts.

2 METHODS

In this paper, we mainly focus on two-channel cDNA arrays, but the main idea can be extended to other types of arrays.

Given a dataset from treatment-control time course design, suppose that it measures the expression levels of n genes at K1 time points/samples under control condition and K2 time points/samples under treatment condition. Let as denote the gene expression levels under the control and treatment condition as $F^{(1)} = (F^{(1)}_{1}, \ldots, F^{(1)}_{n})$ and $F^{(2)} = (F^{(2)}_{1}, \ldots, F^{(2)}_{n})$ correspondingly. In both matrices, each row is the expression levels of a gene across different time points while each column stands for all the n genes’ expression levels at one specific time point.

First, under each condition (treatment or control) and for each gene i, we can define the relationships between gene i and all the other genes by calculating the distance $d(i,j)$ between their expression patterns. Several metrics can be used to describe this distance including the Euclidean distance, Pearson correlation coefficient and etc. Then for gene i, we obtain two distance vectors $d(1) = (d(1)_1, \ldots, d(1)_i, \ldots, d(1)_n)$ and $d(2) = (d(2)_1, \ldots, d(2)_i, \ldots, d(2)_n)$, where $d(1)_i$ and $d(2)_i$ are the distances between the expression patterns of gene i and gene j under the control condition and the treatment condition, respectively.

Second, for each distance vector under each condition, the ranks of all the genes j \neq i are calculated and denoted by $r^{(1)} = (r^{(1)}_1, \ldots, r^{(1)}_i, \ldots, r^{(1)}_n)$ and $r^{(2)} = (r^{(2)}_1, \ldots, r^{(2)}_i, \ldots, r^{(2)}_n)$, where $r^{(1)}_j$ and $r^{(2)}_j$ are the rank of $d^{(1)}_j$ in $d^{(1)}$ and $d^{(2)}_j$ in $d^{(2)}$, respectively. Then the change of the relationships between gene i and gene j under the two conditions can be described as

$$\Delta r_j = |r^{(1)}_j - r^{(2)}_j|,$$

where $j = 1, 2, \ldots, n$ and $j \neq i$.

Third, we define a ‘neighborhood’ for gene i because the change of gene i under the two conditions should not be described by the change in the relationships between gene i and all the other genes. Two types of genes are included in the ‘neighborhood’. The first type include those genes that have very similar expression profiles with gene i because these genes tend to be functionally associated with gene i. However, if we only consider this type of genes, when all the neighbors of gene i are perturbed by the treatment to the same level, we would not see significant change in the ‘neighborhood’ of gene i although gene i does change under the two conditions. To make up this problem, we include the second type of genes into the ‘neighborhood’ of gene i, which have very large distance from gene i under either condition. These distant genes usually consist of genes from various function categories and may have no biological association with gene i. When all the neighbor genes are perturbed at the same level, these distant genes will have large changes in their relationships with gene i because most of them may not be perturbed together with gene i or may be perturbed in very different ways from gene i. With all these considerations, we have the following three definitions of ‘neighborhood’:

(1) q-proximal neighborhood: $G_1^{(q)}(i) = \{j : d^{(1)}_j \leq d^{(1)}_i (1 - q)\}$,

(2) q-distal neighborhood: $G_2^{(q)}(i) = \{j : d^{(1)}_j \geq d^{(1)}_i (1 - q)\}$,

(3) q-two-end neighborhood: $G_3^{(q)}(i) = G_1^{(q)}(i) \cup G_2^{(q)}(i)$,

where $d^{(q)}_j$ is the qth lower quantile of the distance vector $d^{(q)}$ and $k = 1, 2$ for the two conditions. So q describes how many genes are included in the ‘neighborhood’ of one gene. Details about how to determine q can be found in ‘Discussion’.

Finally, given the value of q and following one definition of ‘neighborhood’, the MARD for gene i is defined as

$$M_i(q) = \frac{\sum_{j \in G^{(q)}(i)} \Delta r_j}{\#G^{(q)}(i)},$$

where $G^{(q)}(i) = G_1^{(q)}(i) \cup G_2^{(q)}(i)$ is the union of the two sets of neighborhood genes of gene i under control and treatment condition, $l = 1, 2, 3$ corresponding to the three definitions of ‘neighborhood’ and $\#G^{(q)}(i)$ stands for the total number of genes inside $G^{(q)}(i)$. So for any given gene i and value of q, we can calculate a MARD for each of the three definitions of ‘neighborhood’.

Having the MARD values of all the genes, we can rank the genes in descending order of their MARD values. The larger the MARD value of a gene is, the larger change the gene has under the treatment and control conditions.

3 RESULTS

We test our approach using two treatment-control time-course microarray datasets. In the first dataset, time courses of gene expression in response to Ca2\textsuperscript{+} were measured with and without the FK506 treatment in budding yeast (Yoshimoto et al., 2002). In both time courses, gene expression levels were measured at four well matched time points after Ca2\textsuperscript{+} addition: 15, 30, 45 and 60 min. Therefore, we refer to this dataset as the aligned time course dataset. The other dataset provides the gene expression profiles across the cell cycle of wild-type budding yeast (Spellman et al., 1998) and the Δfkh1Δfkh2 double mutant (Zhu et al., 2000). The two time courses were measured independently by two research groups and different sampling schemes were used. Therefore, it is difficult to directly compare the two time courses. We refer to this dataset as the unaligned time-course.

3.1 MARD analysis of the aligned time course data

3.1.1 Ca2\textsuperscript{+} response w/o FK506 inhibition data  Calcinurin is a Ca2\textsuperscript{+}/calmodulin-dependent protein phosphatase. It is activated by specific environmental conditions, including exposure to Ca2\textsuperscript{+}
or Na\(^+\), and then induces gene expression by regulating the activity of the transcription factor Crzl1/Crz1p. The effects of Ca\(^{2+}\) and Na\(^+\) can be counteracted by FK506, which is an inhibitor of the calcineurin protein, thereby shutting down the entire signaling pathway (Fig. 1). To screen for calcineurin-dependent genes regulated by Ca\(^{2+}\), Yoshimoto et al. (2002) performed four groups of cDNA microarray: (1) Ca\(^{2+}\) time course, (2) Ca\(^{2+}\) + FK506 time course, (3) Ca\(^{2+}\) + FK506/Ca\(^{2+}\) and (4) \(\Delta crz1/crz\text{-}1\); Ca\(^{2+}\). In experiment (1) and (2), yeast samples were collected at \(t = 15, 30, 45\) and \(60\) min after being exposed to Ca\(^{2+}\) and Ca\(^{2+}\) + FK506 separately, and were compared with sample collected at \(t = 0\). In experiment (3), direct comparison was made between samples collected from the FK506-treated and control samples at 15 and 30 min after Ca\(^{2+}\) addition. In experiment (4), direct comparison was made between samples collected from wild-type and \(\Delta crz\text{-}1\) strain at 15 and 30 min after Ca\(^{2+}\) addition. The authors identified 153 calcineurin-dependent genes activated by Ca\(^{2+}\) based on microarray data from all the four experiments.

Our aim is to identify the genes significantly perturbed by the inhibition of calcineurin with FK506. Since FK506 blocks the calcineurin/Crz1p signaling pathway, we would expect that the genes directly related to this pathway are more severely perturbed than other genes. According to our approach, the degree of perturbation of a gene is measured by its neighborhood-change between the treatment and control time-course experiments. We only use the data from experiment (1) and experiment (2) and regard Ca\(^{2+}\) + FK506 time course as treatment and Ca\(^{2+}\) time course as control. Totally there are about 6000 genes whose expression levels were measured in the two time courses. We filter out genes with missing values in either time course after which 5052 genes left. Since genes with constant expression levels across the time points in both time courses are of no interest, we remove 20% constantly expressed genes with the smallest variation across all the time points in the two time courses. For the remaining 4042 genes, we apply the MARD analysis (two-end neighborhood with an informative fraction \(q = 1\%\) in this paper, see ‘Discussion’ for determination of \(q\)). Note here only genes activated by Ca\(^{2+}\) are of interest, we use ratios rather than log transformed ratios as the expression measurements to lower the MARD value repressed genes. Detailed explanation will be given in ‘Discussion’.

### 3.1.2 Identification of perturbed genes

We calculated the MARD values for all the 4042 genes and the distribution of them is shown in Figure 2A. Biologically speaking, after the inhibition of calcineurin by FK506, we would expect dramatic neighborhood changes for genes that are directly related to Calcineurin/Crz1p signaling pathway. Therefore these genes are expected to have high MARD values. On the other hand, housekeeping genes, which are essential for cell survival, tend to be less severely affected by any perturbation, since significant change in the activities of these genes may be lethal to Yeast. Consequently, these housekeeping genes should have lower MARD values. As shown in Figure 2A, the histogram of MARD shows a notable heavy tail on the right-hand side and a small peak on the left-hand side, which seem to be the calcineurin/ Crz1p pathway related genes and housekeeping genes, respectively. We investigate those genes with small MARD values (on the left hand), it turns out that most of them are housekeeping genes, such as ribosomal protein genes. Certainly, those genes with large MARD values (on the right side) are more of interest. We validate their association with calcineurin/ Crz1p pathway through comparing with previous studies (Yoshimoto et al., 2002; Cyert, 2003).

### 3.1.3 Consistency with previous study

We checked the consistency of our identified genes with those identified as differentially expressed in Yoshimoto et al. (2002). Yoshimoto et al. applied a two-step analysis to identify calcineurin dependent genes activated by Ca\(^{2+}\). First, they selected 934 Ca\(^{2+}\)-activated genes that were induced more than 2-fold at either 15 or 30 min after Ca\(^{2+}\) addition in experiment (1). Second, they assessed the extent to which the expression of each of these genes was reduced by calcineurin inhibition using direct comparison of FK506-treated and non-FK506-treated cells exposed to Ca\(^{2+}\) in experiment (3). Genes identified by this analysis are based on direct ratio measurements (Ca\(^{2+}\) addition 15 or 30 min’ versus ‘Ca\(^{2+}\) addition 0 min’ and ‘FK506-treated’ versus ‘non-FK506-treated’) in both steps and thereby are of high confidence. However, some calcineurin dependent Ca\(^{2+}\) activated genes may be missing, because: (1) They did not take into account the genes that activated by Ca\(^{2+}\) only at 45 or 60 min. (2) The arbitrarily determined 2-fold threshold may filter out some interested genes. Our MARD analysis aims to find the genes that were significantly perturbed in terms of neighborhood by FK506 treatment. We only consider the two time courses in experiments (1) and (2).

Despite the differences between our method and the approach in Yoshimoto et al., the two sets of identified genes are highly consistent with each other. Yoshimoto et al. identified 153 calcineurin dependent Ca\(^{2+}\) activated genes, among which 111 are present in our dataset (4042 genes included in total). To make a fair comparison, we select the top 111 genes with the highest MARD values as listed in Supplementary Table 1. Among these 111 genes, 63 genes are also identified by Yoshimoto et al. with a \(P\)-value of \(5.7 \times 10^{-77}\). The consistency of our result with that of Yoshimoto et al. is better illustrated in Figure 2B. Most of the genes contain the Crz1p binding motif in their promoter regions, suggesting that they were directly regulated by Crz1p. As can be seen from Figure 2B, genes with higher MARD values are more likely to be reported as calcineurin dependent Ca\(^{2+}\) activated genes in (Yoshimoto et al., 2002). Specifically, all the top 27 genes with highest MARD values...
are among the 153 genes identified by Yoshimoto et al. More interestingly, we found that crz1 itself is significantly perturbed by FK506 according to our result (with rank of 95) while it is not identified as calcineurin-dependent gene by Yoshimoto et al. It turns out that crz1 gene encodes an auto-regulated transcription factor, i.e. it regulates the transcription of itself (personal communication with Martha Cyert, Stanford University).

We also apply two-way ANOVA and EDGE analysis for the data (Park et al., 2003; Storey et al., 2005). The two-way ANOVA analysis results in 167 genes whose expressions are significantly different between the FK506 treated and non-FK506 treated time courses with a significance level \( \alpha = 0.01 \). Among these genes only 6 fall into the 153 genes identified by Yoshimoto et al. If we reduce the significance level to 0.05, 783 genes are identified, among which 54 are also within the 153 genes. However, the EDGE program results in no differentially expressed genes between the two time courses with a false discovery rate <10%. This may be caused by the lack of replicates or the small number of time points in the experiment.

### 3.1.4 Consistency with direct comparison

Because the sampling time points are well matched between the treatment and control in this dataset, it is possible to directly calculate the gene expression profile changes between treatment and control. Here, we would expect the neighborhood change of a gene to be consistent with its expression pattern change for the following reasons: (1) the biology system is robust (Li et al., 2004; Alon et al., 1999), only a small fraction of genes have significant expression changes in response to a nonlethal perturbation; (2) we use Euclidean distance to measure the neighborhood of genes. On the other hand, our approach explicitly uses more information about the gene–gene relationship than direct comparison of gene expression patterns, some differences are also expected.

The change of gene expression patterns in treatment and control is defined as the normalized Euclidean distance:

\[
L(Y^{(1)}, Y^{(2)}) = \frac{\|Y^{(1)} - Y^{(2)}\|}{\|Y^{(1)}\| + \|Y^{(2)}\|},
\]

where \( Y^{(1)} \) and \( Y^{(2)} \) are gene expression profiles in treatment and control, respectively.

Fig. 2. MARD analysis of the Ca\(^{2+}\) Response w/o FK506 Inhibition Data. (A) Distribution of the MARD values (informative fraction \( q = 1\% \)) of the 4042 genes. Shown in y-axis is the number of genes as a function of the MARD value shown in x-axis. Threshold at the vertical dash line result in 142 genes. (B) The ranks of the genes identified by previous studies. Bars below the thick line: identified genes in (Yoshimoto et al., 2002); bars above the line: genes with known functions (Cyert, 2003). (C) Consistency between our approach and direct comparison method. Shown in y-axis is the MARD value and shown in x-axis is the normalized Euclidean distance [Equation (1)]. The identified genes in (Yoshimoto et al., 2002) are marked as black stars. The Pearson correlation coefficient between MARD and normalized Euclidean distance is 0.8436. (D) Relationship between MARD values (\( q = 1\% \)) and lethality. Shown in y-axis is the percentage of lethal genes with the corresponding logarithm-transformed MARD values shown in x-axis.
where $Y^{(1)}_g$, $Y^{(2)}_g$ are the two time courses of gene $g$ under control and treatment conditions, respectively and $||\cdot||$ is the $L_2$ norm in this study.

We plot the MARD value of each gene versus the expression pattern change for each gene in Figure 2C. As we can see from the plot, genes with higher MARD values tend to have larger expression pattern change in treatment versus control time course. The correlation coefficient between the MARD values and the normalized Euclidean distances of all the genes is 0.844. Furthermore, most of the genes identified by Yoshimoto et al. have higher MARD values than the normalized Euclidean distances. This indicates that the MARD-score based analysis has a higher discriminant power.

### 3.1.5 Essentiality and MARD

Due to the robustness of a biological system, we would expect small neighborhood changes in response to a perturbation for genes that are essential for cell survival. Therefore, we studied the relationship between the MARD value and essentiality of genes. Systematic gene deletion experiments have been performed in yeast (Winzeler et al., 1999). In total, 5860 yeast genes are deleted and 1117 (19%) of them are identified as essential genes which means that single deletion of these 1117 genes is lethal for cells grow in YPD medium.

We rank the MARD values for all the 4042 genes, and calculate the lethality rate using ranked from $i$ to $i + 100$ for different $i = 1, 10, 20, \ldots$. The lethality rate is defined as the fraction of essential genes in the gene set. We plot the MARD values against the resulting lethality rate for each gene set in Figure 2D. As shown in the figure, the lethality rate decreases from 56 to 5% with the increase of MARD values. This is reasonable because the lethality rate describes how essential the genes in the gene set are to the organism. If most of the genes inside a gene set are essential, the perturbation by the treatment on them should be relatively small because significant perturbation on them may be lethal to the organism. Now we have smaller MARD values for more essential gene set. This means that our MARD statistic is a good measure of the effect of the treatment on each gene. Since our method actually measures the change in the neighborhood of each gene, this also supports our rationale that genes, which are more severely affected by a treatment, tend to have larger neighborhood changes and thereby higher MARD values. In addition, these results also show that gene relationship network is robust, because essential genes that play important roles tend to be less affected by a treatment.

![Fig. 3. MARD analysis of the wt/Δfkh1Δfkh2 cell cycle data (Spellman et al., 1998; Zhu et al., 2000). (A) Distribution of the MARD value (informative fraction $q = 1\%$) of the 5525 genes. (B) Relationship between MARD value and lethality. Shown in y-axis is the percentage of lethal genes with the corresponding logarithm-transformed MARD values shown in x-axis.](https://academic.oup.com/bioinformatics/article-abstract/22/21/2650/251710)

### 3.2 MARD analysis of the un-aligned time course data

Fkh1 and Fkh2 are two yeast transcription factors involved in cell cycle regulation. Deletion of each of them may cause mis-regulation of some genes, especially cell-cycle related genes. Spellman et al. performed a time-course experiment to identify cell-cycle regulated genes in wild-type yeast (Spellman et al., 1998). Zhu et al. (2000) performed another time-course experiment in which fkh1 and fkh2 were knocked-out. Two clusters of genes (CLB2 and SIC1) that show different expression patterns in the Δfkh1Δfkh2 mutant were identified as Fkh1 or Fkh2 dependent genes by Zhu et al. Since the two time course datasets have different sampling schemes, the expression patterns of genes in them cannot be directly compared. Bar-Joseph et al. (2003) identified 30 cell-cycle genes and 22 non-cycling genes as differentially expressed by representing expression patterns of genes by function curves and comparing directly the function curves.

### 3.2.1 Identification of perturbed genes

After filtering out the genes with more than one missing values, we calculate MARD values (two-end neighborhood with an informative fraction $q = 1\%$) for the remaining 5525 genes to identify genes significantly perturbed by Δfkh1Δfkh2 knockout. In this dataset we use log transformed ratios as the expression measurements instead of using the ratios directly. The reason for doing this can be found in Discussion. The distribution of MARD values of all the 5525 genes is shown in Figure 3A. We selected the top 100 genes with the highest MARD values which are listed in Supplementary Table 2. Among these 100 genes, 41 genes are cell-cycle related genes according to the result from Spellman et al. ($P$-value = $4.9 \times 10^{-14}$). While comparing these 100 genes with the results by Bar-Joseph et al. (2003), 13 genes show up in their top 30 cycling genes and 9 genes show up in their top 22 non-cycling genes ($P$-value $= 5.7 \times 10^{-11}$). Finally when comparing our results with that by Zhu et al. (2000), we find that none of the up-regulated genes and 7 ($P$-value $= 7.2 \times 10^{-3}$) of the down-regulated genes identified by them are in our top 100 genes. Again a negative correlation between MARD value and essentiality is observed which is shown in Figure 3B.

Δfkh1Δfkh2 double mutation have global effects on cell growth. With this double mutation, the cells show pseudohyphal and invasive growth, unusual cell morphology, and slow growth rates (Zhu et al., 2000). Consistent with these phenotypes, many of the
top 100 genes identified by our approach are involved in cell-cycle, cell wall organization, amino acid synthesis or pseudohyphal growth. For example, MEP1 (with rank of 76) is an ammonium permease that regulates pseudohyphal differentiation in response to ammonium limitation (Lorenz and Heitman, 1998). TEC1 (with rank of 43) is a transcription factor which is involved in pseudohyphal growth (Kohler et al., 2002). This gene is also identified by Zhu et al. but not by Bar et al. (Zhu et al., 2000). PCL2 (CLN4, rank 24) is a G1 cyclin, which associates with Pho85p cyclin-dependent kinase (CDK) to contribute entry into the mitotic cell-cycle and is essential for cell morphogenesis (Measday et al., 1994; Moffat and Andrews, 2004). We also checked the genome-wide binding data (Lee et al., 2002) that described the association of Fkh1p and Fkh2p with genes expressed in G1 and S phases, and found 7 genes bound by Fkh1p (P-value = 0.014) and 15 genes bound by Fkh2p (P-value = 4.3 × 10^{-6}) in the top 100 perturbed genes.

4 DISCUSSION
4.1 Measurement selection
As mentioned, the measurement for gene expression is ratio in the first dataset (Ca^{2+} response w/o FK506 inhibition), while the measurement is log transformed ratio in the second dataset (Δfkh1Δfkh2/wtcellcycle). In the first data, we want to identify calcineurin-dependent Ca^{2+} activated genes as done by Yoshimoto et al. Although there do exist some genes that are repressed by Ca^{2+}, we are more interested in Ca^{2+} activated genes as what Yoshimoto et al. did in their paper. So in order to make our results comparable with Yoshimoto’s results, we reduce the influences of Ca^{2+} repressed genes by using ratio rather than log ratio as the measurement for gene expression. In such a situation, the expression ratios for Ca^{2+} repressed genes are limited to [0,1], while expression ratios for Ca^{2+} activated genes are always greater than 1. Since we use Euclidean distance in calculating the change in neighborhood, the genes identified by MARD analysis tend to be genes that activated by Ca^{2+} in either FK506 treated or non-treated time courses, and most Ca^{2+} repressed genes are ignored. We note that this is a special case, in most cases we want to treat gene activation and repression equivalently and therefore log ratio should be used as the gene expression measurement.

4.2 Neighborhood selection
To identify genes that are differentially expressed between treatment and control time courses, we construct gene relationship networks for the the two time courses, respectively. The genes substantially affected by the treatment are expected to show dramatic changes in its neighbor genes. Essentially, here the neighbor genes refer in particular to those genes that have small Euclidean distances with a specific gene, namely, proximal neighbor genes. However if only proximal neighborhood changes are considered, one may fail to identify some co-affected genes because most of their proximal neighbor genes are similarly affected by the treatment and therefore there is no notable change in the neighborhood. To make up this problem, we take advantage of those distant neighbor genes which have the largest Euclidean distances from the specific gene. We note that there is no underlying biological relationship between these distant neighbor genes. However, the distant neighbors of a gene tend to be from various function categories and widely distributed in the relationship network. So the change in the relationships between these distant neighbor genes and the specific gene may imply the global position change of the gene in the whole relationship network, which cannot be captured by proximal neighborhood change. As shown in Figure 4, we studied the effectiveness of the three neighborhood definitions (proximal only, distal only and two-end neighborhood definition) by setting informative fraction q to range from 0.1 to 50%. MARD analysis is performed on the Ca^{2+} response time courses with and without FK506 treatment. For each setting of the informative fraction q and neighborhood definition, the number of those genes that are among the top 142 genes in our result and also among those calcineurin dependent Ca^{2+} regulated genes reported by Yoshimoto et al. (2002) is calculated to measure the effectiveness of each neighborhood definition. The result shows that the distal neighborhood definition can achieve more effectiveness if a large q (>4%) is used. But when q is small, the effectiveness of the distal neighborhood definition is much worse than the other two definitions. In comparison, the proximal neighborhood or two-end neighborhood can achieve good effectiveness across a wide range of q. According to our experience of MARD analysis in various datasets, including two datasets not reported in this article, we suggest using both-end neighborhood definition. Our general strategy of selecting the fraction value q is as follows: first, we try MARD analysis for q in a range, say [0.008, 0.05] as used in the above cases; second, we look for a stable set of genes that is invariant across the range of q values; third, we validate the function of these genes by scientific facts reported in the literature; fourth, we make further hypotheses based on the computational results. This strategy works well in the examples we have analyzed so far. We hope this bioinformatic methodology will benefit other researchers. We note that the informative fraction q for proximal and distal neighborhood do not necessarily need to be equal in the two-end neighborhood. Further improvement is expected by setting different values for them.
4.3 Metric selection
The relationship between genes can also be measured by other metrics besides the Euclidean distance. For example, Pearson correlation is often used to measure the similarity between expression profiles of genes, based on which gene co-expression networks are constructed and used to predict gene functions, infer transcriptional regulatory networks, and so on (van Noort et al., 2003; Segal et al., 2003). In addition, comparing correlations between genes across experiments has been proposed to further improve these studies (Zhou et al., 2005). Generally, the correlation can be applied to infer the gene relationship network in MARD analysis. But in practice, there are some disadvantages for using correlation as the metric. First, most of the time course data contain only a small number (<10) of time points, therefore it is inappropriate to use correlation to measure the gene relationship. Second, several works have shown that co-expression network is scale free in topology (van Noort et al., 2004; Barabási and Oltvai, 2004), and the number of nodes with a given degree follows a power law distribution. In contrast to ‘random’ networks, scale-free networks are highly non-uniform. In the gene co-expression networks, the hub genes have many co-expressed neighbors, while most other genes have only a few neighbors. This feature may be taken into account when correlation is used for the MARD analysis and some revisions may be required.

4.4 Significance level of MARD values
As so far, all the analysis and results show that the MARD statistic does reflect the degree of perturbation of genes by a treatment. A higher MARD value implies a more severe perturbation. However, it is difficult to assign a significance level to an observed MARD value because MARD values for all the genes are strongly dependent with each other. For example, if a gene is substantially affected by a treatment, the MARD values of its neighbor genes will also tend to be large. In addition, it is hard to perform permutation analysis for time courses as used in SAM (Tusher et al., 2001). In a static microarray experiment, one permutes samples to ‘balance’ the case and control datasets and thereby estimate the false discovery rate based on the ‘balanced’ datasets. But in time course data, different time points provide different aspects of gene expression. Therefore it is inappropriate to permute the time points to calculate the significance level of MARD for each gene.

5 CONCLUSION
We have developed a new method to identify differentially expressed genes between treatment and control time courses. Rather than comparing gene expression patterns in the two time courses directly, we construct gene relationship networks for each of the time courses and then measure the neighborhood change of each gene in the two networks. The genes that are substantially affected by the treatment, i.e. differentially expressed genes, are those that have a remarkable neighborhood changes.

We applied our method to both aligned and un-aligned time course datasets. The results in the aligned dataset show that (1) genes with high MARD values exhibit different expression levels between treatment and control time course in all or a subset of time points; (2) the genes identified by our method are consistent with previous studies, where additional well-designed experiments are performed to ensure the accuracy of the result; (3) we also found some genes that are related to the pathway of interest but failed to be identified by previous approaches. Our method avoids direct comparison of gene expression patterns between time courses, therefore it is insensitive to sampling effect. We do not require equal or ‘aligned’ sampling time points in the treatment and control time courses. So our method can be used to compare time courses from different sources as shown in the unaligned w/\Delta\Delta/\Deltah/\Deltah/2 cell-cycle dataset. In addition, the MARD value can roughly reflect the importance of a gene in the cell system. Genes with small MARD values tend to be house-keeping genes, most of which are essential for cell survival.

ACKNOWLEDGEMENT
The authors thank Martha Cyert for providing us with information on the regulation of Crz1. This work is supported by the grant R01 GM75308-01 from NIH. This work is partially supported by Center of Excellence in Genome Science at University of Southern California, the NIH P50 HG002790 grant.

Conflict of Interest: none declared.

REFERENCES