BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i23–i30
DOI: 10.1093/bioinformatics/bth915

Deconvolving cell cycle expression data with complementary information

Ziv Bar-Joseph[1,*], Shlomit Farkash[2], David K. Gifford[3], Itamar Simon[2] and Roni Rosenfeld[1]

[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA, [2]Hebrew University Medical School, Hadassah Ein Kerem, Jerusalem, 91120, Israel and [3]MIT CSAIL, 200 Technology Square, Cambridge, MA 02139, USA

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: In the study of many systems, cells are first synchronized so that a large population of cells exhibit similar behavior. While synchronization can usually be achieved for a short duration, after a while cells begin to lose their synchronization. Synchronization loss is a continuous process and so the observed value in a population of cells for a gene at time t is actually a convolution of its values in an interval around t. Deconvolving the observed values from a mixed population will allow us to obtain better models for these systems and to accurately detect the genes that participate in these systems.

Results: We present an algorithm which combines budding index and gene expression data to deconvolve expression profiles. Using the budding index data we first fit a synchronization loss model for the cell cycle system. Our deconvolution algorithm uses this loss model and can also use information from co-expressed genes, making it more robust against noise and missing values. Using expression and budding data for yeast we show that our algorithm is able to reconstruct a more accurate representation when compared with the observed values. In addition, using the deconvolved profiles we are able to correctly identify 15% more cycling genes when compared to a set identified using the observed values.

Availability: Matlab implementation can be downloaded from the supporting website http://www.cs.cmu.edu/~zivbj/decon/decon.html

Contact: zivbj@cs.cmu.edu

INTRODUCTION

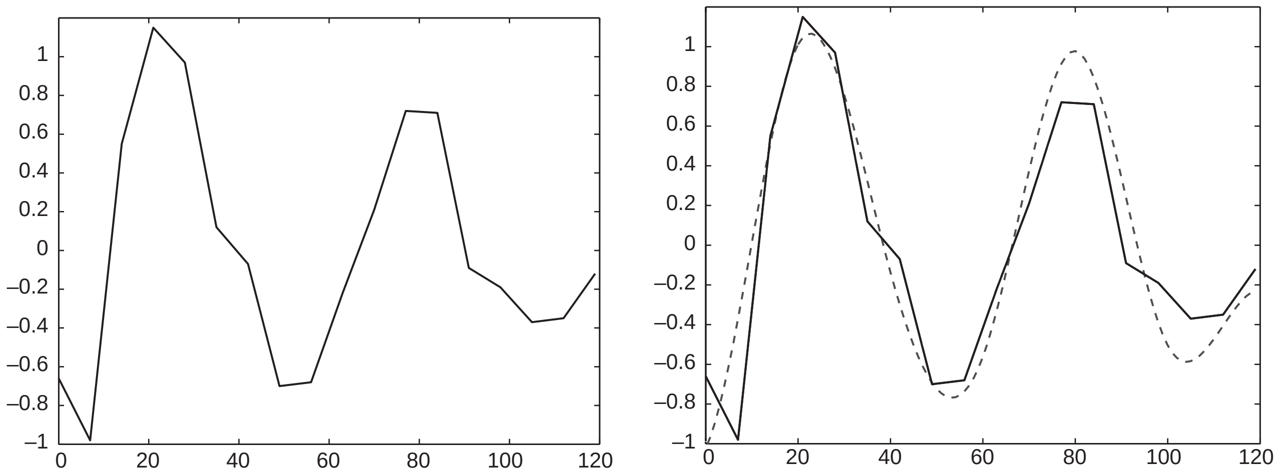
Cyclic systems, such as the the cell cycle (Spellman *et al.*, 1998) and circadian clock (Panda *et al.*, 2002) play a key role in many biological processes, including development and cancer. Due to our inability to profile single cells, expression experiments that study these systems are usually carried out by synchronizing a population of cells. Synchronization is achieved by first arresting cells at a specific point and then releasing cells from the arrest so that at the beginning of the experiment all cells are at the same point.

Even with the best synchronization method cells do not remain synchronized forever. For yeast, cells seem to remain relatively synchronized for two cycles (Spellman *et al.*, 1998; Shedden *et al.*, 2002B) while wild type human cells lose their synchronization very early (Shedden *et al.*, 2002A) or halfway through the first cycle (Whitfiled *et al.*, 2002) depending on the arrest method. Synchronization loss is a continuous process. Even for yeast, cells are much less synchronized during the second cycle when compared with the first cycle. This causes the peak expression value to be lower in the second cycle and the lowest expression value to be higher for most cycling genes (Fig. 1). Thus, the expression value measured for a gene g at time t is actually a convolution of the true expression values of g at an interval around t. Deconvolving the measured expression values to more accurately represent single-cell behavior will allow us to improve the results of algorithms that generate models for the cell cycle system. In addition, the deconvolved profiles improve our ability to identify cycling genes in yeast, and may lead to the discovery of a similar set of cycling genes in humans.

While we can indirectly detect loss of synchronization using expression data, there are two other methods that are more suitable for this task: fluorescence-activated cell sorting (FACS) analysis and budding index. FACS is a method for determining the DNA content of individual cells. Cells are inserted into a narrow tube, and at the end of the tube the DNA content of each cell is measured using a laser reader. Budding index is the process in which cells are counted under the microscope to determine the presence and size (small or large) of buds for each cell. While these methods have been used to validate synchronization experiments, both can only assign cells into one of three cell cycle phases: G_1, S and G_2/M[1]. While this data is useful for determining the rate

[*]To whom correspondence should be addressed.

[1] FACS can actually determine a distribution for cells in the S phase as well, however, this data is relatively noisy and many researchers use only the total amount of cells in S. See, e.g. Whitfiled *et al.* (2002)



(a) Observed values

(b) Deconvolved values.

Fig. 1. Agreement between first and second cycle for Smc3, one of the 800 cycling yeast genes. Using the observed values, there is a difference between the peak and lowest expression value when comparing the first and second cycle. These differences are drastically reduced when deconvolving the measured expression data using the algorithm discussed in this paper. See also the Results section.

of synchronization loss, it cannot be directly used to reconstruct expression profiles. The main problem is that this data is too coarse, partitioning cells into only three phases while expression profiles are continuous in nature.

In this paper, we present a method for deconvolving population effects by combining budding index or FACS data with gene expression data. We assume that following release from arrest, each cell proceeds according to its own internal clock. Clock speeds for all cells are assumed to be normally distributed with mean 1 (the real time) and an unknown variance. The biological basis for this model is the observation that cells are growing at (slightly) different rates which in turn affects their entrance into S phase and progression through the rest of the cell cycle. In order to test the validity of our model we generated new budding index data for yeast. As we show in Results section, our model fits the observed data very well, even though it contains far fewer parameters than data points.

In order to deconvolve the measured expression data we need to assign a continuous representation to each gene. Due to noise and missing values, interpolating individual genes does not work very well. Instead, a method that uses co-expressed genes to constrain spline assignment to individual genes was presented in Bar-Joseph *et al.* (2002). Here we modify this method to deconvolve expression values as well. The resulting profiles are the single-cell expression values for each gene, and these profiles allow us to correctly identify cycling genes that cannot be identified when relying on the measured values, or an interpolated version of these values.

Related work

Fluorescence-activated cell sorting and budding index were used in the past to validate synchronization in gene expression experiments. For example, Spellman *et al.* (1998) used both methods to validate that yeast cells can be synchronized using a variety of arrest methods. Whitfiled *et al.* (2002) used FACS data to show that unlike wild type cells, human cancer cells remain relatively well synchronized for two cycles. However, in all previous work on gene expression data, these methods were not used to determine a synchronization loss model, as we do in this paper. We are not aware of papers that used these data sources to deconvolve expression data.

Determining the rate of synchronization loss was addressed previously in the biological literature, though not in the context of expression data. Creanor and Mitchison (1994) presented a heuristic method which relies on cell division time to determine this rate. Unlike their method, our algorithm can also use the rate in which cells progress from G_1 to S and from S to G_2/M, leading to a more accurate model. Further, unlike our algorithm, their method is not model based and requires manual adjustments.

Shedden and Cooper (2002B) used a Fourier analysis algorithm to test the synchronizations of different arrest methods. Wichert *et al.* (2004) presented methods for identifying periodically expressed genes using statistical methods. While these methods can be used to detect synchronization loss, they cannot be directly used to deconvolve expression profiles as we do in this paper.

Lu *et al.* (2003) presented a method for deconvolving static expression data in yeast. Their goal is to model the expression values of genes in steady state as a linear combination of different cell cycle phases. Unlike our method, their method assumes a set of perfectly synchronized expression values, and cannot be directly used to deconvolve time series expression data. In addition, their method relies solely on the expression data, and thus cannot be used in organisms in which cells cannot be synchronized beyond one cycle (such as humans).

Zhao *et al.* (2001) assigned pre-determined curves (sinusoids) to yeast expression profiles. By relying on the accuracy of the first cycle they were able to reconstruct a better representation for the total expression profile of each gene and detect a better set of cycling genes. While their method is useful, unlike our method it cannot be extended to other organisms since it relies on the presence of a synchronized first cycle. In addition, since it only relies on expression data, this method needs to make a very strong assumption about the shape of the curve. In contrast, by using additional information (FACS and budding index) our algorithm works for any type of curve.

While this paper was under review, Lu *et al.* (2004) presented a different method for resynchronizing time series expression data. As we do in this paper they assume that cell cycle rates for yeast cell population follow a normal distribution. Their method fits a convolved sinusoid to time series expression data. Unlike our method, their method relies solely on the measured expression data, requiring the existence of a strong cyclic signal for all cell cycle genes. Unfortunately, such strong cyclic signal does not exist for many other organisms, including human fibroblast cells. As mentioned above, such cells are not synchronized for even one cycle. In contrast, by relying on external measurements (such as FACS data) our algorithm can generate a synchronization loss model for any cell type, and use this model to deconvolve cell cycle expression.

METHODS

Budding index data

Spellman *et al.* (1998) performed budding index analysis in conjunction with their expression profiling but this data was reportedly lost (Spellman and Sherlock, personal communication). We have thus performed additional budding index analysis. Yeast cells (W303 strain Z1321) were grown to OD600 of 0.2 in YPD. The cells were synchronized by adding alpha factor (5 μg/ml) to the growth medium. After 2 h incubation the culture were completely arrested in G_1 (all the cell were without buds). The synchronization was released by washing out the alpha factor from the medium (by pelleting the cells and changing to a fresh medium) and the cells were grown for additional 90 min. Samples were taken every 15 min, fixed (1% formaldehyde) and observed under a light microscope. For each time point 200 cells were counted and the fraction of cells with no bud, a small bud (smaller than one half of the yeast cell) or a large bud was documented. In Figure 2 we present the results of one of these experiments. In that figure we annotated the no bud fraction as G_1, small bud as S and large bud as G_2/M. See supporting website (http://www.cs.cmu.edu/~zivbj/decon/decon.html) for complete results.

Modeling synchronization loss

Let t denote universal or external time. We will assume that each cell has its own 'internal clock' which controls its progression through the cell cycle. Each cell has an intrinsic speed v, and therefore its own 'internal time', vt.

We will further assume that the speeds of the internal clocks in the cell population follow a Gaussian (Normal) distribution[2]. Based on the observed budding index data, it is clear that this speed is restricted to a limited range (Fig. 2). The following describes our assumption and observed restrictions for v:

$$v \sim N(\mu, \sigma) \quad 0.5 \leq v \leq 1.5.$$

We define the average speed to be $\mu = 1$ (so that clocks are distributed around the real observed time). We are thus left with a single parameter, σ, for characterizing cell cycle rate variation.

At time $t = 0$, all cells transition from the arrested state into G_1. Subsequently, cells with higher internal speed are the first to transition into the next phase. Let d_G, d_S, d_M denote the duration of the G_1, S and G_2/M phases, respectively, in cells with $v = 1$. For notational convenience, let $d_{CYC} = d_G + d_S + d_M$.

Let $m_G(t)$ denote the fraction of cells that are in G_1 at time t, and similarly define m_S and m_M for cells in S and G_2/M. Under our model, the cells found in the G_1 phase at time t are those whose speed v obeys $0 \leq vt < d_G$ (first cycle) or $d_{CYC} \leq vt < d_{CYC} + d_G$ (second cycle). This is true as long as no cell was able to start the third cycle, i.e. $t \leq 2 \cdot (d_{CYC})/v_{max}$, which is the case for the data analyzed in this paper. Therefore, the fraction $m_G(t)$ of cells in phase G_1 is:

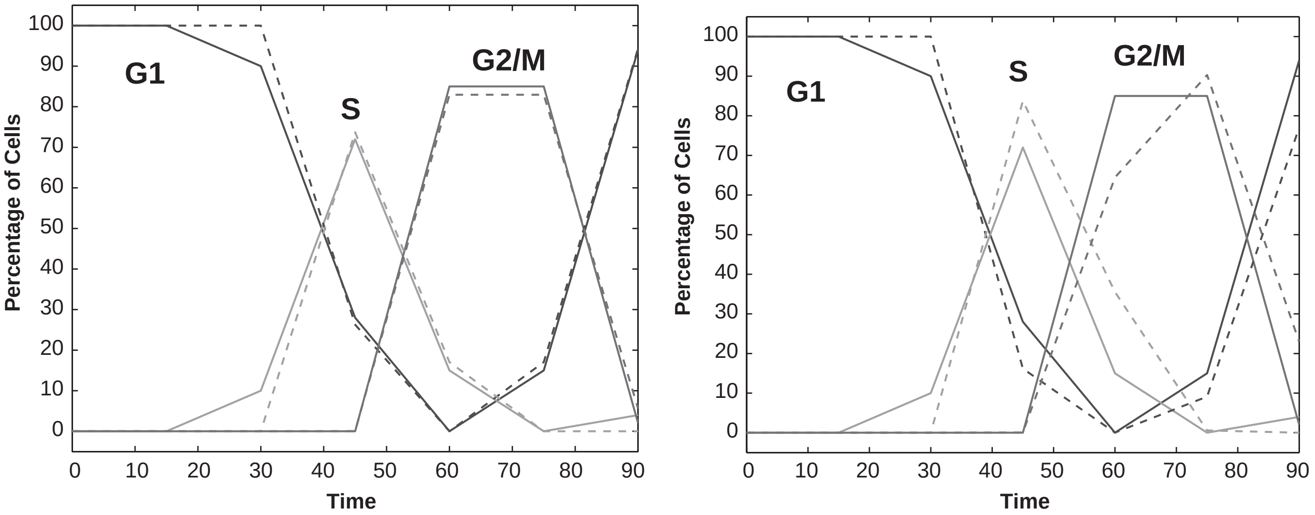
$$m_G(t) \propto \int_0^{tv=d_G} e^{(v-1)^2/2\sigma^2} dv + \int_{tv=d_{CYC}}^{tv=d_{CYC}+d_G} e^{(v-1)^2/2\sigma^2} dv.$$

And similarly for $m_S(t)$ and $m_M(t)$. Note that we have ignored the normalization constants because they can be recomputed by requiring $m_G + m_S + m_M = 1.0$ for all t.

Let $V_G(t), V_S(t), V_M(t)$ be the empirical measurements (e.g. FACS or budding index) of phase proportions at time t, and let $(t_1, t_2, \ldots, t_k)$ be the times at which they were taken. We can now fit the parameters of our model (σ, d_G, d_S, d_M) by minimizing the sum squared difference between the predicted and observed values, namely

$$\begin{aligned}\mathrm{ERR}(\sigma, d_G, d_S, d_M) = \sum_{i=1}^{k} [&(m_G(t_i) - V_G(t_i))^2 \\ &+ (m_S(t_i) - V_S(t_i))^2 \\ &+ (m_M(t_i) - V_M(t_i))^2].\end{aligned}$$

[2] An alternative assumption would have been a Poisson distribution of phase lengths. Biologically our model is more reasonable, at least for G_1, because it says that cells exit G_1 after going through a stochastic growth process, rather than by some random spontaneous event.



(a) Fitting one dataset

(b) Fitting all three datasets

Fig. 2. Comparison between observed budding index values (solid line) and reconstructed values (dashed lines). (**a**) Comparing measured and fitted values based on one experiment. (**b**) Comparing fitted values determined using the three repeats to measured values of the same experiment shown in (a). In both cases the agreement is quite good, despite the fact that budding data is noisy and the fitted model has many fewer parameters than the number of observed points. This indicates that the synchronization loss model we assume can be used to explain synchronization loss in cell cycle experiments.

Since the error function to be minimized is not obviously convex, more than one local minimum may exist. Nonetheless, we can efficiently find a good local minimum using a successive line minimization algorithm such as Powell's Method (Press *et al.*, 1992, p. 412), which starts from an arbitrary point in parameter space and cycles through the parameters, finding a minimum along one dimension at a time. This algorithm is guaranteed to converge (albeit to a local minimum), and in practice often converges rapidly. As we show in the Results section, for budding index data the resulting parameters fit the measured data very well.

The method described above can be modified to fit a more general model of synchronization loss, in which a different Gaussian (a distinct σ) is used for each phase. This will increase the number of variables to six (two additional variance terms are required). Interestingly, even though this model is more complex, for the budding index and expression data analyzed in this paper the model discussed above (using the same σ for all phases) seems to give the best results.

Deconvolving time series expression data

We present a method for deconvolving gene expression data using the synchronization loss model discussed above. Note that such a deconvolution cannot be performed without interpolating the observed values. Even if we knew the phase distribution of cells at time t, because of the relatively low sampling rates, we cannot obtain the values for the interval around t without a continuous representation for the measured data.

Several methods have been suggested to interpolate time series gene expression data. In prior work (Bar-Joseph *et al.*, 2002, 2003) we have shown that cubic spline interpolation, where spline assignment to individual gene is constrained by co-expressed genes, outperforms other interpolation methods for such data. Here we present an extension to this method allowing it to deconvolve expression data for individual genes while still relying on co-expressed genes to counter noise and missing value.

Deconvolving expression data using splines The observed value for gene i at time t is actually a convolution of is expression values at an interval around t. Let $u_i(t)$ represent the underlying expression value for i at time t, and set

$$g(x,t) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{(x/t-1)^2/2\sigma^2},$$

where σ^2 is the variance determined for the synchronization loss. $g(x,t)$ represents the fraction of cells that are at time x when the real time is t. Let $Y_i(t)$ be the observed value for i at time t. Then we can write:

$$Y_i(t) = \int_0^\infty u_i(x) g(x,t) dx + \epsilon. \tag{1}$$

That is, $Y_i(t)$ is a convolution of is expression values (u_i) where the weighting is based on the percentage of cells that are at time x when the real (experiment) time is t. ϵ is a noise term which is assumed to be normally distributed with mean 0.

We use cubic splines to represent u_i. Cubic splines are a set of piecewise cubic polynomials, and are frequently used for fitting time-series and other noisy data. Specifically, we use B-splines, which can be described as a linear combination of a set of basis polynomials. By knowing the value of these splines at a set of control points, one can generate the entire set of polynomials from these basis functions. Due to noise and missing values, fitting splines to individual genes leads to overfitting of the expression data. Instead, we use mixed effects models, which combine gene specific and class information to constrain spline assignment using co-expressed genes. Using such a model, u_i can be written as

$$u_i(t) = s(t)(\mu_k + \gamma_i). \tag{2}$$

Here, μ_k is the mean spline control point for class C_k (the class to which i belongs) and γ_i is the gene-specific spline control point. The parameters of this model are determined using an EM algorithm. In the E step, we determine class membership for each gene and the other parameters of the model are maximized w.r.t. the class assignment in the M step. See Bar-Joseph *et al*. (2002) for complete details.

We now use our continuous spline representation for u_i to deconvolve the measured expression values. Substituting Equations (2) into (1) we get

$$Y_i(t) = \int_0^\infty s(x)(\mu_k + \gamma_i)g(x,t)dx + \epsilon, \tag{3}$$

$$= \sum_j (\mu_{k,j} + \gamma_{i,j}) \int_0^\infty s_j(x)g(x,t)dx + \epsilon, \tag{4}$$

where $\mu_{k,j}$ and $\gamma_{i,j}$ are the j-th entry in the class mean and gene-specific control points, respectively, and $s_j(x)$ is the j-th entry of the spline coefficients evaluated at time x. Set

$$b_j(t) = \int_0^\infty s_j(x)g(x,t)dx.$$

We can then write

$$Y_i(t) = b(t)(\mu_k + \gamma_i) + \epsilon, \tag{5}$$

where the j-th entry in $b(t)$ is $b_j(t)$.

Equation (5) replaces the spline coefficients $s(t)$ from Equation (2) with a weighted spline coefficients $b(t)$, where the weighting is determined using our synchronization loss model. However, apart for this difference (and the noise we assume for measurement error), the two equations are the same. Note that since σ^2 has been fixed, $b(t)$ does not contain any parameters, and can be computed using numerical integration. Thus, we can use the same EM algorithm mentioned above to fit the parameters of the mixed-effects models, and deconvolve the measured expression data by replacing every occurrence of $s(t)$ with the corresponding $b(t)$. Due to lack of space we do not repeat the details of this algorithm. The reader is referred to (Bar-Joseph *et al*., 2002) for more details.

RESULTS

We have tested our algorithm using budding index and gene expression data from yeast cells. The main reason we have used yeast is because unlike other organisms, yeast is relatively synchronized for two cycles. This, and the fact that a lot is known about cycling genes in yeast allows us to validate the results of our algorithm, as we discuss below. In addition, the fact that so many researchers have used yeast cell cycle expression data to model networks in the cell makes the reconstruction of the true underlying single-cell profiles an important goal.

Modeling synchronization loss

We have repeated the budding index analysis three times, and have used the algorithm discussed in the Methods section to determine the rate of synchronization loss. Overall, our model fitted the data very well. Figure 2 shows the observed and reconstructed values when fitting our model to one of the experiments and to all three. Note that, although our algorithm uses only four parameters to fit 21 (for one experiment) or 63 (for three) observed points, the fit is very good, indicating that our model can be used to explain synchronization loss in yeast cells.

Using the complete set of experiments we have determined that the mean duration of the cell cycle (d_{cyc}), is 84 min (this value ranged from 80 to 88 min for the individual experiments, indicating a good agreement between repeated measurements). The G_1 phase was determined to be 41 min, S phase 17 min and G_2/M 26 min.

The SD of the internal clocks (σ) ranged between 0.07 and 0.11 for the individual experiments. Combining the three repeats resulted in $\sigma = 0.09$ which is the value we used for deconvolving expression data.

Deconvolving yeast cell cycle expression data

We have used alpha synchronized expression data (Spellman *et al*., 1998) to test our algorithm. This data contained 18 time points sampled uniformly every 7 min between 0 and 119. The duration of the cell cycle was shorter (65 min) for the expression data, perhaps because of differences in the time of arrest between the budding and expression experiments (Spellman, personal communication). Since cells progress quicker in the expression experiment we have slightly scaled our estimation of σ accordingly, and set $\sigma^2/0.09^2 = 84/65 \Rightarrow \sigma = 0.1$.

Below we present a comparison of the results of our deconvolution algorithm (DECON) with the observed values (VALUES) and with an interpolated version of these values based on mixed effects models (MEFFECTS).

Comparing first and second cycles In order to test the resulting deconvolution, we have looked at the ability of our algorithm to improve the agreement between the first and second cycle. For this, we have used the 800 cycling yeast genes determined by Spellman *et al*. (1998). Note

Table 1. Global comparison between peak and low expression differences for the three datasets

	Diff. peak	Diff. low
VALUES	0.14	0.25
MEFFECTS	0.09	0.20
DECON	0.09	0.11

Note that for DECON, both differences are small and are within the measured noise range (0.11). This indicates that the synchronization loss model we inferred from the budding index data agrees well with the measured expression data.

that our deconvolution method does not rely on the relationship between the first and second cycle, and so even though our algorithm uses a more complex model this test is valid. First, we have compared the difference between the peak and bottom points of the first cycle and the corresponding points in the second cycle for these genes. Table 1 presents the average square difference between these points for VALUES, MEFFECTS and DECON. For both peak and bottom, DECON and MEFFECTS did better than VALUES, because of their ability to overcome noise and missing values. For peak points, both DECON and MEFFECTS performed well, and the differences were within the range of the measured noise variance (0.11). However, for the bottom point, DECON did much better than MEFFECTS. While the difference using DECON was within the range of the measured variance, the MEFFECTS result was almost twice that much. The reason for the difference between peak and bottom values for MEFFECTS might be because a large proportion of the cycling genes are in G1. G1 genes peak early, and reach their bottom values toward the end of the cycle. Thus, these genes are more synchronized in their second peak compared to their second bottom. We have also performed a more global test, by aligning the two cycles using MEFFECTS and DECON and computing the resulting alignment error for all genes. These results too confirmed that the reconstructed curve achieves better agreement between the two cycles [results are omitted due to lack of space, see supporting website (http://www.cs.cmu.edu/~zivbj/decon/decon.html) for details]. In Figure 3, we present plots of the average expression profiles for genes in two of the cell cycle phases. Note that in both cases, the reconstructed curves result in a better agreement between the first and second cycle.

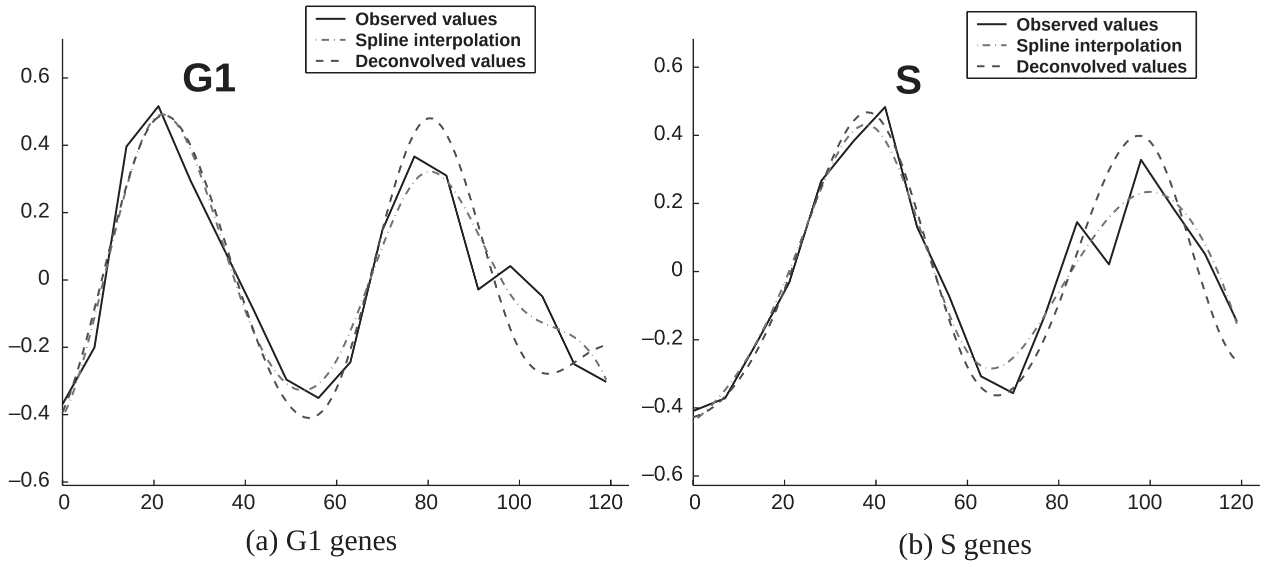
Identifying cell cycle genes We have tested whether our deconvolution results can help in identifying cell cycle genes that cannot be identified using the observed values alone. To this end we have a used the Fourier Proportion of VariancE (PVE) method developed by Shedden and Cooper (2002B) which compares the ability of periodic and a-periodic curves to explain the expression profile. Note that for identifying cycling genes we cannot rely on the 800 cell cycle genes from Spellman *et al.* (1998) since this set was determined using the alpha values. Instead, we first complied a list of the top 800 cycling genes using two other cell cycle expression datasets: Cdc15 (24 time points) and Cdc28 (17). The complete list is available from the supporting website (http://www.cs.cmu.edu/~zivbj/decon/decon.html). We denote this list by CYC.

In order to determine a cutoff for cycling genes, we randomized the expression values for each gene, and applied the interpolation and deconvolution algorithms to this data as well. For each set of profiles (VALUES, MEFFECTS and DECON) we determined the PVE threshold that detected only 1% (60) of the genes in the random set, and used it to select all genes (from the 6000 yeast genes) that were above this threshold when using the original data. The MEFFECTS result did not distinguish well between the randomized and real data, detecting only 141 genes above the noise level. The main reason MEFFECTS did not perform well for this task is because of the tendency of spline approximation (without deconvolution) to smooth the observed measurements. While this helps in overcoming noise, it also flattens the profiles for both random and real data. On the other hand, both VALUES and DECON were able to detect a large number of genes above the noise level (510 and 529, respectively). We intersected these lists with CYC and found that the list generated from DECON was in much better agreement. VALUES contained 167 genes from CYC (p-value $= 10^{-33}$) while DECON had 195 such genes (p-value $= 0$). Thus our deconvolved profiles were able to correctly identify 15% more genes when compared with the observed values alone, indicating the importance of applying this method to measured cell cycle data.

We also looked at a list 113 genes that were previously detected by Cho *et al.* (1998) as cycling using the cdc28 data only, but were omitted from the 800 cycling genes list by Spellman *et al.* (1998) because they did not seem to cycle in alpha and cdc15 [see supporting website (http://www.cs.cmu.edu/~zivbj/decon/decon.html) for the list]. Using the deconvolved values we found that eight of these genes were determined to be cycling in alpha, even though they cannot be detected using the measured values alone. Figure 4 presents three examples of such genes.

DISCUSSION AND FUTURE WORK

Many systems can only be studied by arresting large population of cells. While arresting cells works well initially, cells lose their synchronization as a result of growth rate and other differences. Thus, the observed values in such studies are actually a convolved version of the underlying single-cell expression values. By deconvolving the observed values we can reconstruct this underlying profile, resulting in an improved ability to detect system-related genes and to model such systems.



(a) G1 genes

(b) S genes

Fig. 3. Comparison between first and second cycle for yeast genes. (**a**) and (**b**) Observed, spline interpolated and deconvolved expression values for genes in G_1 and S phases. Note that in both cases the second peak is correctly higher in the deconvolved profiles, resulting in a better agreement between the first and second cycle.

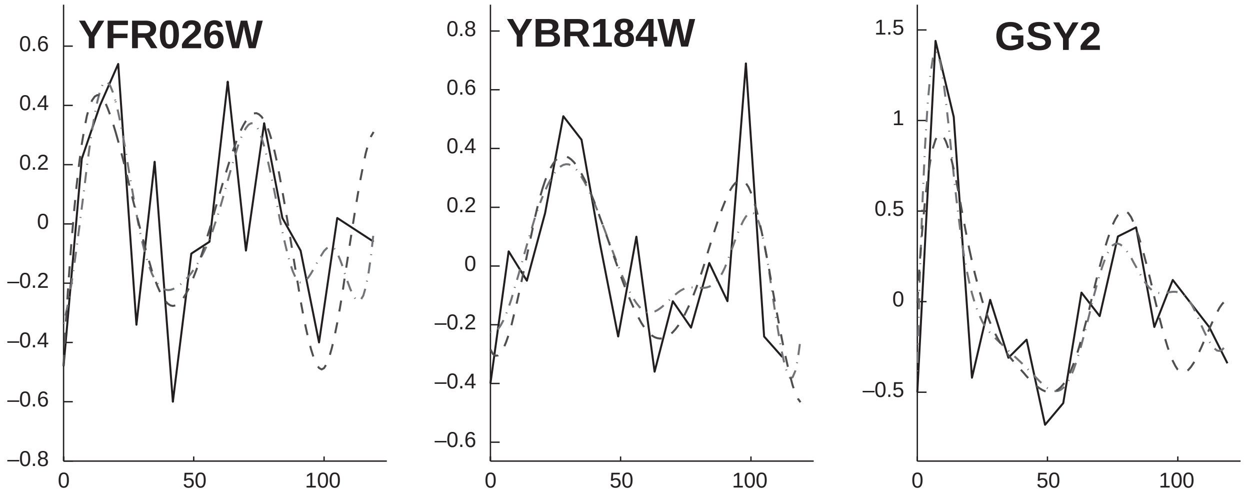


Fig. 4. Observed (solid black line), interpolated (dotted red) and deconvolved (dashed blue) values for three cycling genes that were correctly identified when using the deconvolved profiles, but were not identified by Spellman *et al.* using the measured alpha values. Note that our deconvolution results overcome noise in the data and achieve better agreement between the first and second cycle.

In this work, we presented a synchronization loss model for the cell cycle system. This model was used to deconvolve cell cycle expression data. We have carried out biological experiments to validate our synchronization loss model. Deconvolving yeast expression data using this model resulted in better agreement between first and second cycle and improved our ability to detect cycling genes.

Cell cycle is tightly linked to development and cancer. An important open problem is to determine the list of cycling human genes. Unfortunately, wild type human cells cannot be synchronized for one complete cycle. In future work we intend to use the method presented in this paper to deconvolve human cell cycle expression data, in order to determine the list of cycling human genes. We believe that this list, and the deconvolved expression profiles will aid future work in modeling the cell cycle system in yeast and humans.

ACKNOWLEDGEMENT

We thank P. Spellman and G. Sherlock for their help. We also thank Tommi Jaakkola and Georg Gerber for comments and discussions. R.R. gratefully acknowledges financial support by National Science Foundation ITR grant 0225656.

REFERENCES

Bar-Joseph,Z., Gerber,G., Jaakkola,T.S., Gifford,D.K. and Simon,I. (2002) A new approach to analyzing gene expression time series data. *Proceedings of the Six Annual International Conference on Research in Computational Molecular Biology* (*RECOMB*), pp. 39–48.

Bar-Joseph,Z., Gerber,G., Gifford,D.K., Simon,I. and Jaakkola,T.S. (2003) Comparing the continuous representation of time series expression profiles to identify differentially expressed genes. *PNAS*, **100**, 10146–10151.

Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Galurelian,A.E., Landsman,D., Lockhort,D.J. and Davis,R.W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.*, **2**, 65–73.

Creanor,J. and Mitchison,J.M. (1994) The kinetics of H1 histone kinase activation during the cell cycle of wild-type and wee mutants of the fission yeast. *J. Cell. Sci.*, **107**, 1197–204.

Lu,P., Nakorchevskiy,A. and Marcotte,E.M. (2003) Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proc. Natl Acad. Sci., USA*, **100**, 10370–10375.

Lu,X., Zhang,W., Qin,Z.S., Kwast,K.E. and Liu,J.L. (2004) Statistical resynchronization and Bayesian detection of periodically expressed genes. *Nucleic Acids Res.*, **32**, 447–455.

Panda,S., Antoch,M.P., Miller,B.H., Su,A.I., *et al*. (2002) Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell*, **109**, 307–320.

Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P (1992) *Numerical Recipes in C, The Art of Scientific Computing, 2nd Edn.* Cambridge University Press, Cambridge.

Shedden,K. and Cooper,S. (2002A) Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc. Natl Acad. Sci., USA*, **99**, 4379–4384.

Shedden,K. and Cooper,S. (2002B) Analysis of cell-cycle gene expression in *Saccharomyces cerevisiae* using microarrays and multiple synchronization methods. *Nucleic Acids Res.*, **30**, 2920–2929.

Spellman,P.T., Sherlock,G., Zhang,M.O., Iyer,V.R., Andees,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. *Mol. Biol. Cell*, **9**, 3273–3297

Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M. and Brown,P.O. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors *Mol. Biol. Cell*, **13**, 1977–2000.

Wichert,S., Fokianos,K. and Strimmer,K. (2004) Identifying periodically expressed transcripts in microarray time series data *Bioinformatics* (in press).

Zhao,L.P., Prentice,R. and Breeden,L. (2001) Statistical modeling of large microarray data sets to identify stimulus–response profiles. *Proc. Natl Acad. Sci., USA*, **98**, 5631–5636.