

# Improved Fourier Transform Method for Unsupervised Cell-cycle Regulated Gene Prediction

Karuturi R. Krishna Murthy

Liu Jian Hua

Genome Institute of Singapore

60, Biopolis Street

Singapore 138672

Republic of Singapore

E-mail: {karaturikm, liujh}@gis.a-star.edu.sg

## Abstract

**Motivation:** Cell-cycle regulated gene prediction using microarray time-course measurements of the mRNA expression levels of genes has been used by several researchers. The popularly employed approach is Fourier transform (FT) method in conjunction with the set of known cell-cycle regulated genes. In the absence of training data, fourier transform method is sensitive to noise, additive monotonic component arising from cell population growth and deviation from strict sinusoidal form of expression. Known cell cycle regulated genes may not be available for certain organisms or using them for training may bias the prediction.

**Results:** In this paper we propose an Improved Fourier Transform (IFT) method which takes care of several factors such as monotonic additive component of the cell-cycle expression, irregular or partial-cycle sampling of gene expression. The proposed algorithm does not need any known cell-cycle regulated genes for prediction. Apart from alleviating need for training set, it also removes bias towards genes similar to the training set. We have evaluated the developed method on two publicly available datasets: yeast cell-cycle data and HeLa cell-cycle data. The proposed algorithm has performed competitively on both datasets with that of the supervised fourier transform method used. It outperformed other unsupervised methods such as Partial Least Squares (PLS) and Single Pulse Modeling (SPM). This method is easy to comprehend and implement, and runs faster.

**Software:** <http://giscompute.gis.nus.edu.sg/cdcAnal>

**Keywords:** Fourier transform, Cell cycle, Gene prediction, Microarray.

## 1. Introduction

Microarray time course measurement of genome-wide mRNA expression levels allows genome-wide prediction of cell division cycle (CDC) regulated genes. In each cell division cycle, cells pass through four phases (M-G2-S-G1-M) in the fixed order. Each CDC regulated gene expresses in one of these phases which give rise to the expectation that the CDC regulated genes would exhibit periodic expression if they were studied for more than one cell division cycle. This is the basis for finding genes with oscillating expression in synchronized cell culture to find the CDC regulated genes. Fourier transform method in conjunction with known cell-cycle regulated genes has been employed for prediction of yeast cell-cycle regulated genes (Spellman et al. 1998 [19]) and HeLa cell-cycle regulated genes (Whitefield et al. 2002 [22]). Knowledge of cell-cycle regulated genes may not be available for several organisms or using the limited known cell cycle regulated genes may introduce phase and profile pattern bias in predicting the other CDC regulated genes. Several methods, such as partial least squares (Johansson et al. 2003 [10]), single pulse modeling (Zhao et al. 2001, [25]), k-means clustering (Tavazoie et al. 1999, [21]), QT-clustering (Heyer et al. 1999, [7]), singular value decomposition (Alter et al. 2000 [1], Holter et al. 2000 [9]), correspondence analysis (Fellenberg et al. 2001 [5]), wavelet analysis (Klevecz, 2000 [11]) also have been applied which do not use training set. But, apart from being computationally intensive and difficult to implement, they either reveal a few CDC regulated genes or yield higher false discovery rate.

Fourier transform (FT) method is sensitive to noise and monotonic components in the gene expression pattern which makes it unsuitable for unsupervised cell-cycle regulated gene prediction [16]. Irregular sampling of gene expression and sampling for non integral number of half cycles introduce phase dependent bias in fourier trans-

form estimation of both amplitude (peak) and phase. The bias is introduced due to the violation of orthogonality assumption [14] in Fourier transform. This paper proposes a simple expression profile normalization method to remove monotonic additive components which can take care of even decaying multiplicative factors, corrective procedures to remove phase dependent bias in the estimation of phase and peak of expression, and proposes scoring function which can take care of deviation of expression profile shape from pure sinusoidal. Apart from these, we also proposed a missing value filling strategy and Gaussian smoothing of the expression profiles which enhance the performance of the fourier transform method.

The proposed method has been applied to publicly available yeast CDC data (Spellman et al. 1998) and HeLa CDC data (Whitefield et al. 2002). The new algorithm, called *Improved FT (IFT)*, outperforms the simple *Fourier Transform (FT)* method while performed competitively with those of supervised fourier transform method. It outperforms the unsupervised methods such as Partial Least Squares (Johansson et al. 2003) in terms of false discovery rate and competitive on recall and prediction stability, but outperforms single pulse modeling (Zhao et al. 2001) in terms of recall, false discovery rate as well as the prediction stability on Yeast CDC data. The results for HeLa CDC data are not available for these methods.

Rest of the paper is organized as follows: Section 2 presents the algorithm. Section 3 presents results and discusses its relative performance. Finally, section 4 concludes the paper.

## 2. Methods

The Improved Fourier transform method involves several steps for each expression profile. The first step is to fill the missing values in the expression profile and smoothing them. The next step is to normalize the profile such that the linear additive component in the profile is removed and the multiplicative monotonic factor is compensated for. Fourier transform is calculated for the normalized profile. The estimates of the fourier components (peak and phase) are corrected to remove phase dependent bias of their estimates caused by irregular and incomplete-cycle sampling. Then the corrected estimate of the fourier peak of the profile in conjunction with the estimated phase is used to find mean square error and sign similarity which compensate for ill-fit of the model and non-sinusoidal expression response. These factors are used together to define CDC score for the profile which is used to predict whether a gene is cell-cycle regulated.

Now onwards we follow the following notation:  $F_i$  is the  $\log_2$  relative expression profile of gene  $i$  and  $F_{it}$  is the value of  $F_i$  at time  $t$ ;  $N$  is the number of samples taken for each

gene;  $N_o$  is the number of samples collected for each gene in each cycle;  $T$  is the period of CDC;  $\omega = \frac{2\pi}{T}$ ; and  $\mathfrak{S}$  is the total period for which the samples were taken for each gene.

### 2.1. Filling Missing Values

The missing values in the profile are filled with the following strategy repeatedly till all missing values are filled:

**Case 1:** Both left and right neighbors exist and they have values assigned. In this case, a missing value is replaced by the average of its neighbors.

**Case 2:** Only one of its neighbors has valid value assigned. Then the missing value is replaced by the average of the neighbor's value and the average of the profile which is calculated as the average of the all values in the original profile.

**Case 3:** Both neighbours are either missing or have missing values. In this case, complete the filling of all missing values falling in the above two cases.

### 2.2. Smoothing Profiles

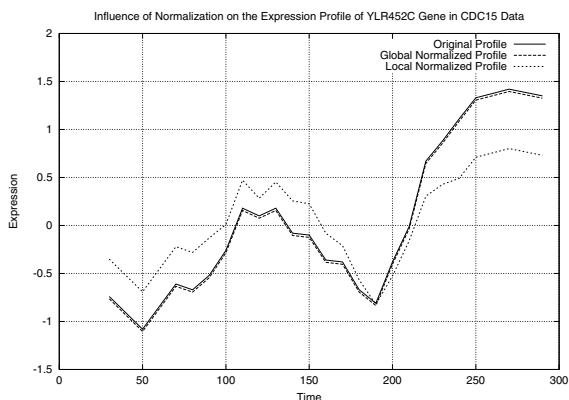
To consider the fact that the microarray measurements of the expression is noisy and may introduce artifactual fluctuations in the expression profiles of genes. These may even be greater than the peak of expression. To deal with these random fluctuations superimposed over the regular periodic expression, the profiles are smoothed according to the following strategy:

$$F_{it} = \sum_x F_{ix} e^{-\frac{(x-t)^2}{2\sigma^2}}$$

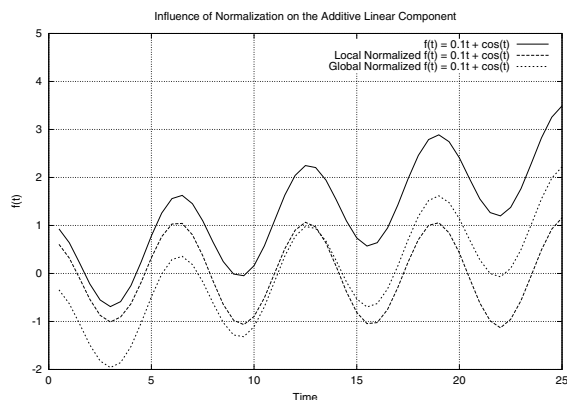
The value of  $\sigma$  was set at  $T/20$ . This signifies the fact that the Gaussian function vanishes almost at  $5 \times \sigma$  from its mean and we expect that the influence of an observation should vanish after quarter cycle away in both directions. This leads to the factor of 20 in the above setting. This setting also has been observed to be optimal on real data.

### 2.3. Zero-mean Local Normalization

Normalizing  $F_i$  such that its mean is zero is common practice. This is usually done by subtracting the mean of  $F_i$  from  $F_{it}$  which we call as *Global normalization*. But the expression profile of a gene may have additive linear component apart from the periodic component (Rifkin and Kim 2002) which has to be removed for better estimate of the fourier transform of the periodic component. This may be achieved by using local normalization by replacing  $F_{it}$  by  $F_{it} - m_{it}$  Where  $m_{it}$  is defined as follows:



**Figure 1. Effect of local normalization on the expression profile of YLR452C gene in CDC15 data.**



**Figure 2. Simulations demonstrating the efficacy of local normalization on additive components.**

$$m_{it} = \begin{cases} \frac{1}{n_t} \sum_{x=0}^T F_{ix} & \text{if } t \in [0, \frac{T}{2}] \\ \frac{1}{n_t} \sum_{x=t-\frac{T}{2}}^{t+\frac{T}{2}} F_{ix} & \text{if } t \in [\frac{T}{2}, \mathfrak{S} - \frac{T}{2}] \\ \frac{1}{n_t} \sum_{x=\mathfrak{S}-T}^{\mathfrak{S}} F_{ix} & \text{if } t > \mathfrak{S} - \frac{T}{2} \end{cases}$$

where  $n_t$  is the number of samples in the chosen window around time  $t$ .

The above formula was derived based on the observation that

$$\frac{1}{T} \int_{x-\frac{T}{2}}^{x+\frac{T}{2}} (y + P \cos(\omega t - \theta) + zt) dt = y + zx$$

The efficacy of the above strategy can be seen in the figures 2, 1 on the simulated expression and YLR452C gene expression. This scheme also helps in compensating for monotonic multiplicative factor  $g(t)$ , mathematically  $g(t)P \cos(\omega t - \theta)$ .  $g(t)$  accounts for slow unraveling of the gene response or dampened gene response which may be expressed as  $e^{\frac{t}{\tau}}$  where  $\tau \in \mathfrak{R}$ . This will be shown in the next section.

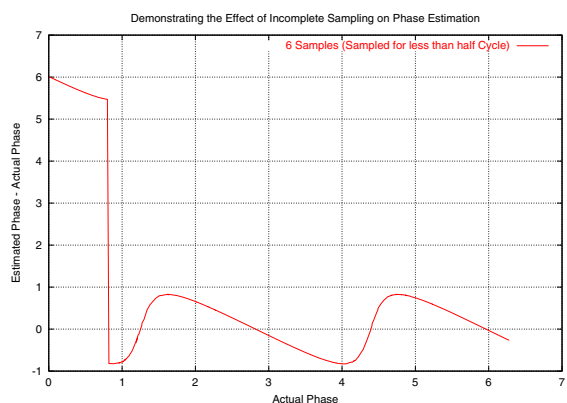
## 2.4. Fourier Transform

The coefficients of the fundamental components of the fourier transform of the expression profile  $F_i$  are obtained as follows:

$$A_i = \frac{1}{N_o} \sum_t \cos(\omega t) F_{it} \quad \text{and} \quad B_i = \frac{1}{N_o} \sum_t \sin(\omega t) F_{it}$$

The peak ( $P_i$ ) and phase ( $\phi_i$ ) of expression are found using the following formulae

$$P_i = \sqrt{A_i^2 + B_i^2} \quad \text{and} \quad \phi_i = \tan^{-1}\left(\frac{B_i}{A_i}\right)$$

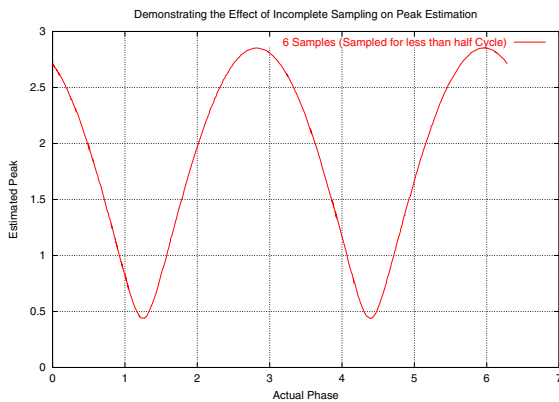


**Figure 3. Simulations demonstrating the variation of  $\phi_i$  with actual phase  $\theta_i$  for the profile  $\cos(t - \theta)$  at the rate of  $N_o = 4\pi$ .**

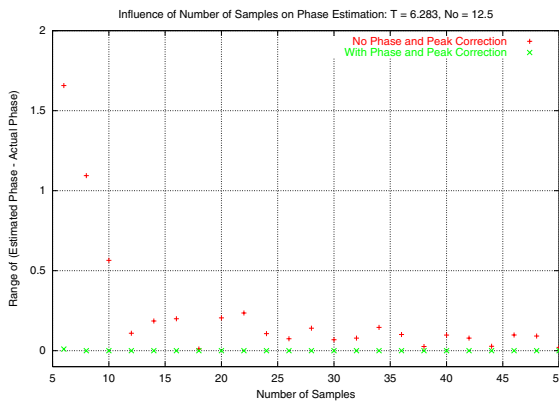
$P_i$  and  $\phi_i$  exhibit variation with the actual phase of  $F_i$  if the number of cycles for which  $F_i$  is sampled is not equal to the integral multiple of half cycles or  $F_i$  were sampled at irregular intervals as shown in figures 3 and 4. The figure 5 shows how much the  $\phi_i$  change as the number of samples are changed from 6 to 50 when number of samples per cycle ( $N_o$ ) is  $4\pi$  and the period of the cycle ( $T$ ) is  $2\pi$  i.e.  $\omega = 1$ .

To correct these artifacts of sampling we used the following table look-up approach.

**2.4.1. Table look-up method:** This method is developed to correct for the artifacts introduced by incomplete and irregular sampling of the expression of genes. In this method, before calculating fourier trans-



**Figure 4. Simulations demonstrating the variation of  $P_i$  with actual phase  $\theta_i$  for the profile generated by sampling  $\cos(t - \theta)$  at the rate of  $N_o = 4\pi$ .**



**Figure 5. Simulations demonstrating the amount of variation of  $\phi_i$  with the number of samples.**

form of the given profiles, we generate several profiles by sampling the function  $\cos(\omega t - \theta)$ , at the times same as that of the sampling times of the given profile, for each value of  $\theta \in [0, 2\pi]$  spaced at regular intervals whose resolution is defined by the user. The fourier transform is estimated for each such profile and table look-up i.e. actual phase vs estimated phase and actual phase vs estimated peak is prepared.

*Correcting for phase:* After finding the phase of expression ( $\phi_i$ ), we search for phase of  $\cos(\omega t - \theta)$  in the table look-up prepared which gives rise to the closest calculated phase if it were sampled at exactly the same time points as that of  $F_i$ . Then the value of  $\phi_i$  is replaced by that of  $\theta$ .

*Correcting for peak:* After having corrected for phase, the calculated peak is divided by the peak obtained for  $\cos(\omega t - \theta)$  if it were sampled at the same time points as that of  $F_i$ .

Now onwards,  $P_i$  and  $\phi_i$  represent corrected peak and phase of expression of profile  $F_i$ . The following subsection defines *CDC score* of the profile  $F_i$ .

## 2.5. CDC Score

Spellman et al. defined CDC score ( $S_i$ ) of a gene  $i$  as a product of estimated peak of expression and peak correlation of  $F_i$  with the profiles generated using the known cell-cycle regulated genes. This strategy was followed by Whitefield et al. also for predicting the Human CDC regulated genes using HeLa cell-line. But in the unsupervised approach, we are not avail of peak correlation. Hence we define two factors, which can substitute the peak correlation, called *mean square error* ( $E_i$ ) and *sign similarity* ( $C_i$ ) for  $F_i$ .  $E_i$  measures how much the model deviates from the original expression profile i.e. how well the best cosine model fits the actual profile and  $C_i$  measures the continuum of agreement of sign of the profile and the cosine model.  $C_i$  is important to take care of noise as well as deviation of shape from sinusoidal and the multiplying factors in the profile. This is motivated by its utility in defining friendly neighbors in [12].  $E_i$  and  $C_i$  are defined as follows:

$$E_i = \frac{1}{N} \sum_t (F_{it} - P_i \cos(\omega t - \phi_i))^2$$

$$M_{ik} = \begin{cases} 1 & \text{if } F_{ik} \cos(\omega k - \phi_i) > 0, \\ 0 & \text{otherwise } \forall i, k \end{cases}$$

$$W_{i0} = 0 \quad \forall i$$

$$W_{i(k+1)} = \max(W_{ik} + \Delta(2M_{i(k+1)} - 1), 0)$$

$$C_i = \sum_{k=1}^S M_{ik} (1 + W_{i(k-1)})$$

Now we are in a position to define CDC score of profile  $F_i$ . It is defined as follows:

$$\text{Score}(S_i) = P_i C_i^\alpha E_i^{-\beta} \quad \text{where } \alpha, \beta \in \mathbb{R}^+$$

$\alpha$  and  $\beta$  were set to 1.0 and 4.0 respectively and  $\Delta$  was set to 0.1 based on the optimum performance on Yeast data. This setting has been observed to be optimum on Hela cell cycle data also.

## 3. Results and Discussion

The *improved FT* method has been applied on both simulated data and real data (Yeast and HeLa CDC data) to reveal its relative merits as compared to the regular *unsupervised fourier transform (FT)* method. The results on the real datasets were compared with the results obtained using supervised fourier transform method. Following the popular

line of thought, we also adopt the strategy of ranking the profiles in the descending order of the CDC scores assigned to them.

To evaluate the algorithms we used *Recall* and *False Discovery rate (FDR)* as criterion. Recall is defined as the fraction of the known cell-cycle regulated genes have their profiles present in the set of predicted CDC profiles. FDR is defined as the ratio of the number of randomly generated profiles whose CDC score is higher than the minimum CDC score among all predicted CDC profiles and the number of profiles predicted to be CDC regulated. We calculate FDR in two ways depending on the distribution used for their generation: (1) Gene randomization in which one random profile is generated for each actual profile by following its distribution; and (2) Complete randomization in which the random profiles are generated following the distribution of the values in the entire CDC data. Similarly, p-value of a gene is fraction of randomly generated profiles have score better than or same as that of the gene under consideration.

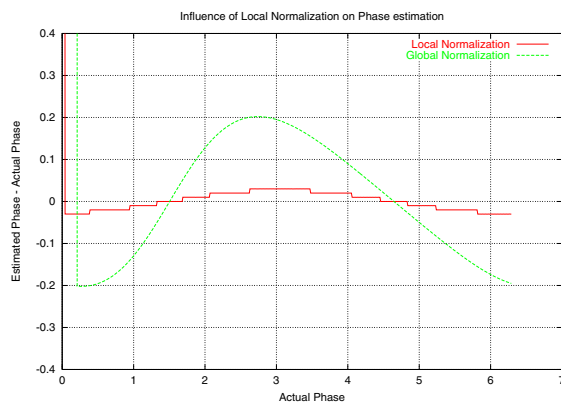
Yeast CDC data is combination of three microarray datasets generated using three different cell-synchronization procedures. Similarly, HeLa is also composed of measurements carried under five different synchronization conditions. The final CDC score evaluation procedure is as follows: (1) each dataset for each organism is individually analyzed and CDC scores are assigned to each profile; (2) final score of a gene is the phasor sum of the weighted CDC scores obtained on all datasets, the weights correspond to the quality of the dataset which is proportional to the number of samples collected in that dataset. In every profile, any expression value different by three fold in the opposite direction of its neighboring values is replaced by missing value. All profiles whose total number of missing values is more than half of observations are assigned CDC score of zero.

### 3.1. Simulated Data Analysis

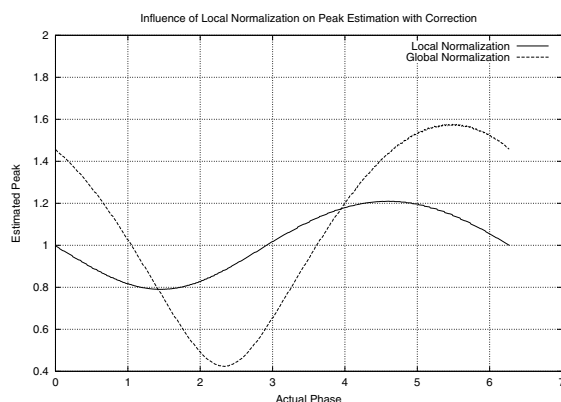
We evaluated the *Improved FT* method on artificial data to assess the merit of the local normalization in *Improved FT* as compared to its FT counterpart and the overall merit.

*Influence of local normalization:* The influence of local normalization on peak and phase estimation of the profiles with additive linear component has been evaluated using the profiles generated by sampling the function  $0.1t + \cos(t - \theta)$  with  $N_o = 4\pi$  for 50 samples i.e. for about 4 cycles.

The plots in figures 6 and 7 reveal that local normalization has significant impact on the estimation of the actual phase ( $\theta$ ) and actual peak which is 1. The plot in figure 8 shows how the CDC score estimated could be improved by using local normalization.



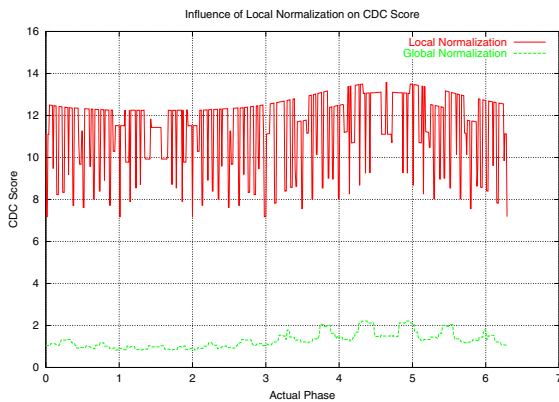
**Figure 6. Comparing local and global normalization on the estimation of  $\theta$  of the function  $0.1t + \cos(t - \theta)$  sampled at the rate of  $N_o = 4\pi$ .**



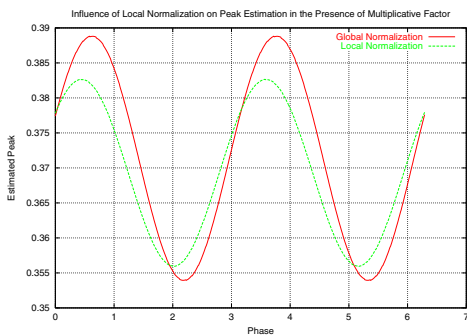
**Figure 7. Comparing local and global normalization on the estimation of peak ( $=1$ ) of the function  $0.1t + \cos(t - \theta)$  sampled at the rate of  $N_o = 4\pi$ .**

The influence of local normalization on peak and CDC score estimation on the profiles with multiplicative factors have been demonstrated in figures 9 and 10 respectively. The function used to generate the profiles is  $e^{-\frac{t}{\tau}} \cos(t)$  with  $N_o = 4\pi$  and  $\tau = 10$  for 50 samples. The conclusions are similar to the additive component case. But the absolute values of peak has been heavily (1/3) underestimated by both methods because of the dampened profile.

*Overall evaluation:* The Improved FT and FT were evaluated using the data generated from the function  $mt + pP \cos(t - \theta) + \epsilon$  where  $m, p, \theta$  and  $\epsilon$  were drawn from  $U(0, 0.1)$ ,  $U(0.1, 1)$ ,  $U(0, 2\pi)$  and  $G(0, 0.25^2)$  respectively.  $U(a, b)$  is uniform probability density function



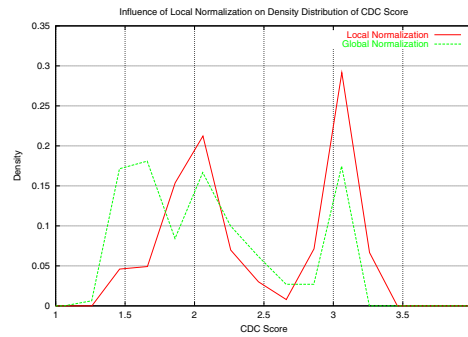
**Figure 8. Comparing local and global normalization on the estimation of CDC Score of the function  $0.1t + \cos(t - \theta)$ , where  $N_o = 4\pi$ .**



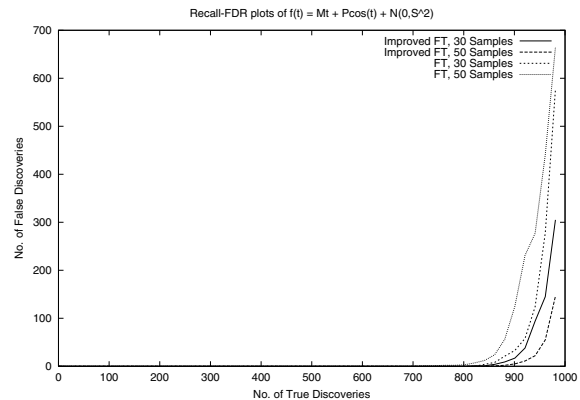
**Figure 9. Comparing local and global normalization on the estimation of peak of the function  $e^{\frac{t}{10}} \cos(t - \theta)$ , where  $N_o = 4\pi$ .**

over the interval  $[a, b]$   $a, b \in \mathfrak{R}$  and  $G(\mu, \sigma^2)$  is Gaussian probability density function with mean  $\mu$  and standard deviation  $\sigma$ . We generated 1000 profiles with  $P = 1$  and 1000 profiles with  $P = 0$  for varying length of sampling with  $N_o = 4\pi$ . The figure 11 show the results of application of Improved FT and FT methods.

The plots from figure 11 clearly show that the Improved FT method is consistently better than FT method and its performance has improved as the number of samples were increased. The fall of performance of FT method as the number of samples increased from 30 to 50 could be ascribed to the fact that the magnitude of the linear component has become more significant making the FT method vulnerable whereas the Improved FT method will become more and more effective as the number of samples were increased.



**Figure 10. Comparing local and global normalization on the distribution of CDC Score of the function  $e^{\frac{t}{10}} \cos(t - \theta)$ , where  $N_o = 4\pi$ .**



**Figure 11. Comparing FT and Improved FT methods on the simulated datasets generated using function  $mt + \cos(t - \theta) + \epsilon$ , where  $N_o = 4\pi$ .**

### 3.2. Yeast CDC Analysis

Both IFT and FT were applied on the three datasets (CDC28, CDC15 and Alpha factor) of yeast [19]. The phase shift of CDC28 and CDC15 experiments from Alpha factor experiment were 0.848rad and 1.39rad. The tables 1 and 2 show the result of the analysis of the three datasets using Improved FT (IFT), FT and *Partial Least Squares* (PLS) [10] methods at the p-value cut off of 0.005 as well as the results of *Single Pulse Modeling* (SPM)[25]. It reveals the performance superiority of the Improved FT method over other unsupervised methods. IFT method could retrieve significantly higher number of genes with very good recall for a given p-value. Its stability is even comparable to FT and PLS methods as seen from the fraction of the total genes occurred in at least two datasets and in all three datasets. These

Method	IFT	FT	PLS	SPM
$\alpha$ factor	375(69)	221(61)	178(56)	328
CDC28	406(53)	177(33)	151(32)	686
CDC15	375(63)	195(50)	227(56)	399
$\alpha$ & CDC28	145(41)	70(25)	56(24)	119
$\alpha$ & CDC15	189(50)	122(42)	114(41)	130
CDC28 & CDC15	140(33)	67(20)	61(20)	147
All 3DS	92(30)	52(16)	42(16)	71
atleast 2DS	290(64)	155(55)	147(53)	254(48)
atleast 1DS	774(91)	386(73)	367(75)	1088(71)

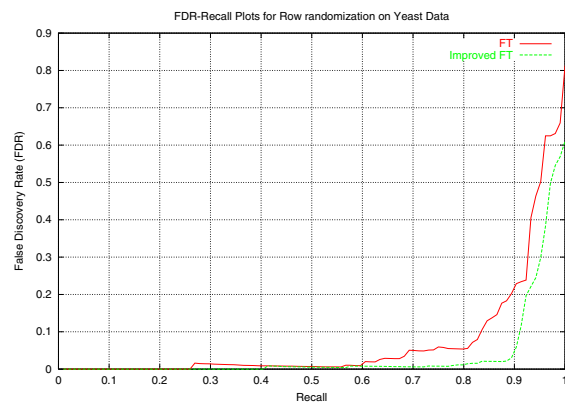
**Table 1. Recall of various methods on Yeast CDC data. Retrieval criterion is p-value  $\leq$  0.005.**

results are significant in the view that the PLS was evaluated using the FDR found for complete randomization which is expected to be lower as compared to the Gene randomization. Apart from this, FDR in PLS was found independent of the original data despite the fact that the PLS is batch analysis i.e. score of one profile depends on the scores of the other profiles. In our study, we have taken the FDR to be of maximum of Gene and complete randomization experiments. It means that our conservative estimates are better than the optimistic estimates of PLS and SPM. Our study even shows that simple FT is outperforming the other complex methods such as PLS and SPM in every respect.

The final CDC score of a gene was obtained by combining the individual CDC scores of the gene as described above. The figures 12 and 13 show the FDR-Recall performance curves for IFT and FT methods on combined CDC

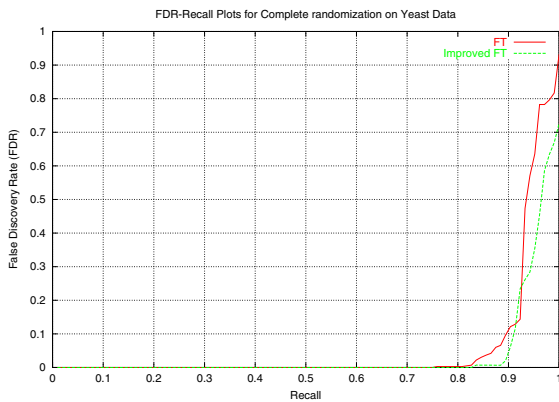
Method	IFT	FT	PLS	SPM
No. of Genes Retrieved	774	386	367	1088
Atleast 2DS	38.0%	40.0%	40.0%	23.0%
All 3DS	11.5%	13.5%	11.0%	6.5%

**Table 2. Prediction stability analysis of various methods on Yeast CDC data. Retrieval criterion is p-value  $\leq$  0.005.**

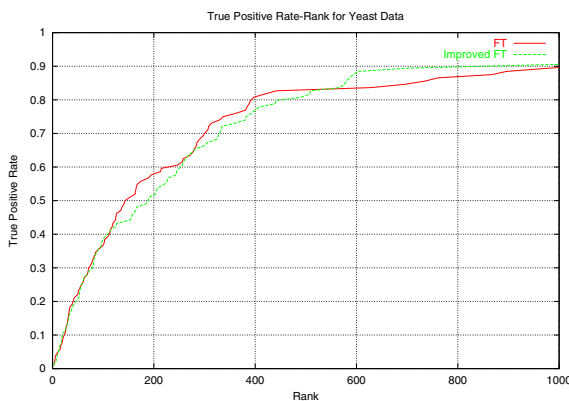


**Figure 12. FDR-Recall plots for IFT and FT methods for Gene randomization on Yeast data.**

scores. They show that the better FDR performance of IFT is noticeable in the last 1/3rd of the top 1000 genes. It means, IFT is required to detect complete list of CDC regulated genes. Figure 14 shows the Recall-Rank plots for FT and IFT. The above observation is true here also. The table 3 shows the result of combining the scores from all three datasets. The results show that the IFT performs competitively with both supervised FT and PLS on recall and outperforms them on FDR both in quantity and its insensitivity to the model for randomization. It clearly outperforms FT in every respect. We have also discovered that IFT has covered 630 of Spellman et al list of 798 genes in the top 800 genes i.e. a overlap of 79%.



**Figure 13. FDR-Recall plots for IFT and FT methods for complete randomization on Yeast data.**



**Figure 14. Recall-Rank plots for IFT and FT methods on Yeast data.**

### 3.3. HeLa CDC Analysis

HeLa CDC data contains five different datasets (TT1, TT2, TT3, TN, TShake) obtained by different synchronization methods. Whitefield et al have eliminated profiles of about 11553 clones which showed combined negative autocorrelation in TT3 and TT2 datasets. For to have fair comparison with the results of Whitefield et al., we also eliminated these profiles from ranking. Whitefield et al have output 1133 clones as showing cyclic variation. The FDR-Recall and Recall-Rank analyses have been carried out on HeLa CDC data obtained from Whitefield et al [22]. The plots in figures 15 and 16 show the FDR-Recall performance curves for IFT and FT methods. Figure 17 shows the relative performance of IFT and FT methods on recall. They also show the similar rel-

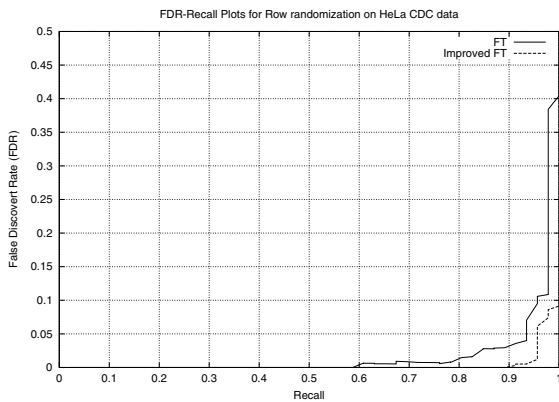
Method	No. of Genes	Recall	False Discovery Rate
SFT	798	95 (91.5%)	3-10%
PLS	755	95 (91.5%)	>10%
IFT	702	93 (89.5%)	2-3%
FT	761	90 (86.5%)	4.2 -14.5%

**Table 3. Comparing the results of CDC analysis using five different methods on Yeast data. The second column shows the rank at which the last true positive has occurred in the top 800 genes. Note that, FDR estimate of PLS is based on complete randomization only.**

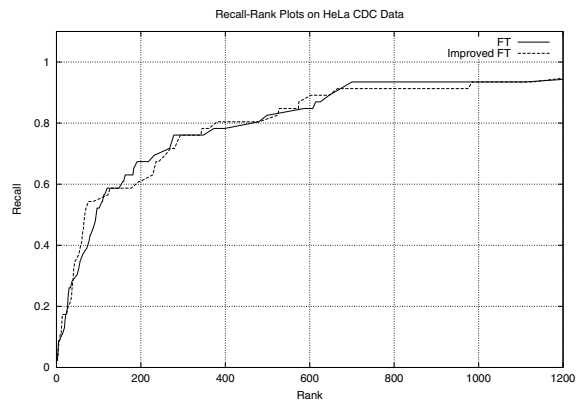
ative performance of IFT over FT as observed in Yeast CDC analysis. The table 4 shows the result of combining the scores from all five datasets for the top 1133 clones.

## 4. Conclusions and Future Directions

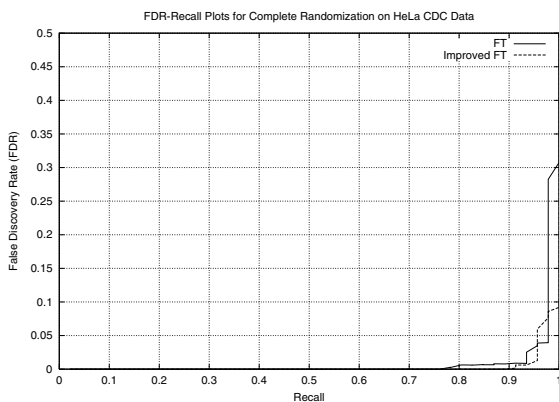
The proposed Improved Fourier Transform method is less sensitive to additive linear component as well as sampling irregularities. Its performance is impressively close to those using prior knowledge while being unbiased in prediction. Its systematic approach and ease of implementation should make it attractive for cell-cycle gene prediction using microarray time course data. We should note that the performance evaluation of the method used by Spellman et al and Whitefield et al is biased towards the training set since the same dataset was used both for training as well as for testing and they belong to certain phases expression. But, our method is unbiased and unsupervised. We have also shown that our algorithm outperformed other recently published algorithms (PLS and SPM) both in terms of prediction stability and FDR-Recall performance. We also observed that the FDR estimation is highly stable across different randomization methods i.e. Gene and complete as opposed to its high variation exhibited by the other methods.



**Figure 15. FDR-Recall plots for IFT and FT methods for Gene randomization on HeLa data.**



**Figure 17. Recall-Rank plots for IFT and FT methods on HeLa data.**



**Figure 16. FDR-Recall plots for IFT and FT methods for complete randomization on HeLa data.**

Method	Rank of occurrence of 43rd Gene	Recall	False Discovery Rate
SFT	850	44 (95.7%)	0.15-0.75%
IFT	938	44 (95.7%)	0.5-0.6%
FT	700	43 (93.5%)	2.5-7.0%

**Table 4. Comparing the results of CDC analysis using three different methods on HeLa CDC data.**

The future direction includes developing periodogram to predict the presence of oscillating profiles in a given microarray time-course data and the analysis of phase of expression stability across various datasets of given organism. We also recognize that the performance of the IFT algorithm has to be studied on several other datasets belonged to a variety of other organisms such as *Arabidopsis Thaliana* [15].

## 5. Acknowledgements

We are grateful to Philip M. Long and Edison T. Liu for their valuable and timely suggestions during this work.

## References

- [1] Alter, O., Brown, P.O. and Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, 97, 10101-10106.
- [2] Burnham, A.J., MacGregor, J.F. and Viveros, R. (1999) Latent variable multivariate regression modeling. *Chemometrics and Intelligent Laboratory Systems*, 48, 16780.
- [3] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. and Davis, R.W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.*, 2, 65-73.
- [4] Doolin, M.T., Johnson, A.L., Johnston, L.H. and Butler, G. (2001) Overlapping and distinct roles of the duplicated yeast

- transcription factors Ace2p and Swi5p. *Mol. Microbiol.*, 40, 422-432.
- [5] Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A., Hoheisel, J.D. and Vingron, M. (2001) Correspondence analysis applied to microarray data. *Proc. Natl Acad. Sci. USA*, 98, 10781-10786.
- [6] Fodor, S.P., Rava, R.P., Huang, X.C., Pease, A.C., Holmes, C.P. and Adams, C.L. (1993) Multiplexed biochemical assays with biological chips. *Nature*, 364, 555-556.
- [7] Heyer, L.J., Kruglyak, S. and Yooseph, S. (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.*, 9, 1106-1115.
- [8] Ho, Y., Costanzo, M., Moore, L., Kobayashi, R. and Andrews, B.J. (1999) Regulation of transcription at the *Saccharomyces cerevisiae* start transition by Stb1, a Swi6-binding protein. *Mol. Cell. Biol.*, 19, 5267-5278.
- [9] Holter, N.S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J.R. and Fedoroff, N.V. (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl Acad. Sci. USA*, 97, 8409-8414.
- [10] Johansson, D., Lindgren, P. and Berglund, A. A Multivariate Approach Applied to Microarray data for Identification of Genes with Cell-Cycle Coupled Transcription. *Bioinformatics*, 19(4):467-473, 2003.
- [11] Klevecz, R.R. (2000) Dynamic architecture of the yeast cell cycle uncovered by wavelet decomposition of expression microarray data. *Funct. Integr Genomics*, 1, 186-192.
- [12] Krishna Murthy, K.R., Vega, V.B., Friendly Neighbors Method for Unsupervised Determination of Gene Significance in Time-course Microarray Data. in the Proc. of IEEE Symposium on BIBE'2004, Taichung, Taiwan, May 19-21, 2004.
- [13] Lander, E.S. (1999) Array of hope. *Nature Genet.*, 21, 3-4. Martens, H. and Naes, T. (1989) *Multivariate Calibration*. Wiley, Chichester.
- [14] Lathi, B.P., *Signals, Systems and Communication*, John Wiley & Sons, 1965.
- [15] Menges, M., Hennings, L., Gruissem, W., and Murray, J.A.H. Cell Cycle-regulated Gene Expression in *Arabidopsis*, *Journal of Biological Chemistry*, 277(44):41987-42002, November 2002.
- [16] Rifkin, S.A. and Kim, J. Geometry of Gene Expression Dynamics. *Bioinformatics*, 18:1176-1183, 2002.
- [17] Rodriguez-Pena, J.M., Cid, V.J., Arroyo, J. and Nombela, C. (2000) A novel family of cell wall-related proteins regulated differently during the yeast life cycle. *Mol. Cell. Biol.*, 20, 3245-3255.
- [18] Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270, 467-470.
- [19] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Bostein, D., and Futcher, B. (1998) Comprehensive Identification of Cell Cycle-regulated Genes of the *S. cerevisiae* *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell*, 9, 3273-3297.
- [20] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, 96, 2907-2912.
- [21] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, 22, 281-285.
- [22] Whitefield, M.L., Sherlock, G., Saldanha, A.L., Murray, J.L., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O. and Botstein, D. Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors, *Molecular Biology of the Cell*, 13:1977-2000, June 2002.
- [23] Wold, H. (1966) Estimation of principal components and related models by iterative least squares. In Krishnaiah, P.A.R. (ed.), *In Multivariate Analysis*. Academic press, New York.
- [24] Wold, S., Ruhe, A., Wold, H. and Dunn, III, W.J. (1984) The collinearity problem in linear regression. the partial least squares approach to generalized inverses. *SIAM J. Sci. Stat. Comput.*, 5, 735-743.
- [25] 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.