THE NATIONAL UNIVERSITY
of SINGAPORE

School of Computing
Lower Kent Ridge Road, Singapore 119260

TRB9/06

Comparison Of Missing Value Estimation In Cell-Cycleregulated Genes Prediction

Guoliang LI, Tze-Yun LEONG and Louxin ZHANG

September 2006

Technical Report

Foreword

This technical report contains a research paper, development or tutorial article, which has been submitted for publication in a journal or for consideration by the commissioning organization. The report represents the ideas of its author, and should not be taken as the official views of the School or the University. Any discussion of the content of the report should be sent to the author, at the address shown on the cover.

JAFFAR, Joxan
Dean of School

COMPARISON OF MISSING VALUE ESTIMATION IN CELL-CYCLE-REGULATED GENES PREDICTION

GUOLIANG LI

School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 117543

TZE-YUN LEONG

School of Computing, National University of Singapore, 3 Science Drive 2, Singapore 117543

LOUXIN ZHANG

Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543

Missing values in the microarray data are problematic in many biological applications. In literature, different methods have been proposed to estimate the missing values, and their performances are evaluated with the normalized root mean square error (NRMSE) on the simulated missing data. Although NRMSE is indicative on the performance of the proposed methods, it does not tell us the methods' effects on the real application. However, in general, the NRMSE in different papers can not be compared directly, since the simulated missing data in the different papers are different. In this paper, we examined six different missing value estimation methods on the real application of cell-cycle-regulated genes prediction, as well as on the simulated missing data. The experiments show that, in terms of NRMSE, our improved knn-based method performs better than the naïve knn-based method, and BPCA and LLSimpute have the smallest NRMSE. In the real application, most of the methods performed similarly in terms of the accuracy of the cell-cycle-regulated genes predicted. Surprisingly, the simple row-mean method can achieve the accuracy as good as other sophisticated methods. And the LLSimpute method performs quite worse in terms of the accuracy of the cell-cycle-regulated genes prediction probably due to overfitting. The results of LLSimpute suggest that the performances in NRMSE and in the real application are not directly correlated. Hence, in the research of the missing value estimation, we need to compare the methods' performance not only in terms of NRMSE, but also in terms of the possible criteria in the real application. The data sets, the improved knn-based method, and results in our experiments are available online http://www.comp.nus.edu.sg/~ligl/missing_values/.

1 Introduction

Missing values in the microarray data are problematic to the biological applications, since most of the current available algorithms require the data to be complete. To address the missing value problem in the microarray data, many methods have been proposed in literature [5,6,9,12,14-16] to estimate the missing values. Among the proposed methods, three representative methods are KNNimpute [14], BPCA [9], and LLSimpute [6]. KNNimpute was proposed by Troyanskaya et al [14], which is probably the first to systematically examine the missing value problem in the microarray data. The authors tried row-mean method, k-nearest-neighbor-based method (KNNimpute) and singular-value-decomposition-based method (SVDimpute). Their experiments showed that KNNimpute is better in terms of the normalized root mean squared error (NRMSE). Oba et al [9] proposed a Bayesian method (BPCA) to fill in the missing values in an

Expectation-Maximization-like repetitive strategy. Their experiments showed that BPCA is better than KNNimpute and SVDimpute in terms of NRMSE. Kim et al [6] proposed a local least square method (LLSimpute) which can take advantage of the properties of k nearest neighbors and linear regression to estimate the missing values. The results from their experiments showed an improved NRMSE compared with KNNimpute and BPCA.

In all the methods mentioned above, the evaluation of the methods is based on NRMSE. NRMSE is a good measure how well the estimated values fit the original values. However, it does not tell us the impact of the methods on the real applications. As we know, missing value estimation is only one of the data pre-processing steps. In the real situations, we have to apply the methods to a real data without knowing the original values of the missings. In such situations, NRMSE can not be applied. To examine the effects on the methods in the real application, we focus on one microarray application – cell-cycle-regulated genes prediction. In literature, there are quite a few researches on the cell-cycle-regulated genes prediction from microarray data [1,2,4,7,8,10,13]. However, the effects of the missing value estimation on the accuracy of the cell-cycle-regulated gene prediction are not well examined.

In this paper, we will examine the effects of six different missing value estimation methods on the accuracy of the cell-cycle-regulated genes prediction, as well as in terms of NRMSE. The experiments show that, in terms of NRMSE, our improved knn-based method is better than the naïve knn-based method; BPCA and LLSimpute have the smallest NRMSE. In terms of the accuracy of the cell-cycle-regulated genes predicted, the experiments show that most of the methods perform similarly. Especially, the simple row-mean method can achieve the similar accuracy as other sophisticated methods. In all the methods, LLSimpute is an exception. It can achieve very good NRMSE, but very bad accuracy on the cell-cycle-regulated gene prediction. It means that the performances on NRMSE and in the real application are not directly correlated. The results suggest that in the research for the missing value estimation, the performance of a new proposed method can not merely compare with the existing methods in terms of NRMSE; it needs to be compared on the real application. Therefore, in the application, we need to be careful in applying the missing value estimation methods.

2 Cell-cycle-regulated genes prediction

In this section, we describe the method that we use to predict the cell-cycle-regulated genes, which includes data pre-processing steps and cell division cycle (cdc) score calculation. Missing value estimation is the first step in data pre-processing. Our improved knn-based method for missing value estimation is also introduced.

2.1 Missing value estimation and an improved knn-based method

Our cell-cycle-regulated gene prediction method is based on Fourier transformation. It requires that the data are complete. In order to fill in the missing values in the microarray data, we proposed an improved k-nearest-neighbor-based method. In theory, the knn-based method will choose the k nearest neighbors in the available data and estimate the

missing values based on these nearest neighbors. But in implementation there are many variants. Firstly, k nearest neighbors can be chosen only from the genes without missing values. This strategy leads to the following problems: 1) if the genes having missing values are more than the genes without missing values, the complete data may not be representative; 2) the selected neighbors may not be appropriate; and 3) the information from the genes with missing values is not utilized. Secondly, k nearest neighbors are allowed to have some missing values, but the missing values are excluded for distance calculation. In this case, the distances between genes with more missing values can be problematically small.

In order to solve the problems mentioned above, we propose an improved knn-based method to fill in the missing values in an iterative way. The intuition is that the distances are more reliable after the missing values have been filled in. Our proposed method first fills in all the missing values with the row means. Second, the k nearest neighbors of genes with missing values are calculated from the filled-in data, and the missing values are re-estimated with the new neighbors. The second step will repeat until a specified number of iterations or until the data converge. Our method is summarized in Table 1.

Table 1 Procedure for the improved knn-based missing value estimation method

1. Fill in the missing values with the row means in the original data
2. Repeat until a specified number of iterations or converge
3. Fill in the missing values with k nearest neighbors from the latest filled-in data

2.2 Smoothing profiles and detrending

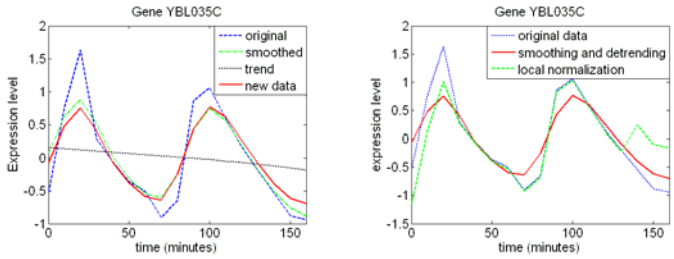
In microarray data, noise as well as missing values may be introduced due to the various reasons, such as the possible artifacts in the experiment process, insufficient resolution, and image corruptions. Also the synchronization method in the gene cell cycle experiments may introduce a stress response from the cell, which can be represented as constant trends in the data. To deal with the noise and the trend, we apply the Gaussian smoothing functions to the microarray data as follows. Equation 1 estimates the smoothed values, Equation 2 estimates the trend values and Equation 3 estimates the new values.

$$S_{gt} = \frac{1}{Z_1} \sum_i x_{gi} e^{\frac{-(i-t)^2}{2\sigma_1^2}} \tag{1}$$

$$T_{gt} = \frac{1}{Z_2} \sum_i x_{gi} e^{\frac{-(i-t)^2}{2\sigma_2^2}} \tag{2}$$

$$X_{gt} = S_{gt} - T_{gt} \tag{3}$$

where the sum is over the different samples of each gene with weights from a Gaussian function, x_{gt} , S_{gt} , T_{gt} and X_{gt} are respectively the original value, the smoothed value, the trend value and the new value of gene *g* at time *t*, Z_1 and Z_2 are normalizing factors to make the coefficients sum to one, σ_1 is equal to the sampling interval and $\sigma_2 = 5\sigma_1$. σ_2 is larger than σ_1 to take more neighbors to calculate the trend. The effect of the smoothing and detrending is shown in Figure 1. Figure 1(a) shows that smoothing filters out the noise in the data, and the detrending minimizes the effect of the constant trend. Moreover, our method estimates the trend continuously which solves the shift problem in the local normalization proposed by Murthy and Liu [8]. The authors mentioned that global normalization by subtracting the mean of X_g from X_{gt} can not deal with the additive linear component from the aperiodic component. They applied local normalization. However, their local normalization needs the period of the cell cycle which is usually not available at this stage of analysis. Even if the period of the cell cycle is available, the local normalization can introduce a shift on the values of the genes at the boundary of the range they defined. There is an example in Figure 1(b). The shift in the value of the gene expression is obvious in the later part of the local normalization.



(a) (b)

Figure 1. Effects of smoothing and detrending. (a) Example for our smoothing and detrending. (b) Example for the local normalization by Murthy and Liu [8]

2.3 Cell cycle period estimation

The cell cycle is decided as the one with the maximum mean cdc score for the known cell cycle genes. In the cell-cycle biological experiments, the researchers sampled the microarray data approximately from two cell cycle periods [10,13]. This information was utilized to estimate the cell cycle period – Assume that the real cell cycles are in the range of 0.7~1.3 times of the observed approximate cell cycles. Then the range is equally sampled with an interval 0.1 minutes and the mean cdc score of the known cell cycle genes is calculated (see Section 2.4). The estimated cell cycle periods for budding yeast are around 92.3 minutes for cdc28, 116.0 for cdc15 and 64.6 for alpha. The estimated cell cycle periods for fission yeast are in the range from 132 to 168 minutes

2.4 Fourier transform and cdc scores

The cdc scores are defined as the magnitudes of gene time series data after Fourier transformation, as shown in the following Equation (4).

$$A = \sum x_t \sin(\omega t + \phi)$$
$$B = \sum x_t \cos(\omega t + \phi) \quad (4)$$
$$C = \sqrt{A^2 + B^2}$$

where x_t are log ratio of the intensities of the expression levels of gene *x* at time *t*, $\omega = 2\pi / T$, T is the cell cycle period, and ϕ is the phase offset. In our work, we set ϕ as 0, since it does not change the magnitude. With the cdc scores, all the genes are sorted in descending order, and the top ones are predicted as cell-cycle-regulated genes. For budding yeast and fission yeast respectively, the cdc scores of the same gene from different experiments are averaged as aggregated scores. The genes are also sorted by the aggregated scores.

3 Missing value estimation in microarray data

3.1 Methods for missing value estimation in our experiments

In our work, we examined six different methods in our experiments.

a) ROWimpute. In this method, the missing values in one row are filled in with the mean of the observed values in the row.
b) SplineImpute. The microarray data for cell cycle genes are time series data. In this method, one cubic spline function is estimated with the observed values in each row and the missing values are estimated as the values of the spline function at the corresponding times.
c) KNNimpute. In this method, the k nearest neighbors are chosen to estimate the missing values [14]. The number k is set to 16 as suggested in the program.
d) IKNNimpute. Refer to Subsection 2.1 for this method. The number k is set to 16.
e) LLSimpute. Kim et al [6] proposed the local least square method for missing value estimation. The optimal number of neighbors is searched in the method.
f) BPCA. Oba et al [9] proposed the Bayesian method to estimate the missing values.

3.2 Evaluation Criteria

In our experiment, different strategies of evaluation are used for validation test. For the simulated missing data, three criteria are used: the root mean square error (RMSE) and two different types of normalized root mean square error (NRMSE), which are RMSE divided by the standard deviation of the real values (denoted as NRMSE1) and by the root mean square of the real values (denoted as NRMSE2) respectively.

For the real dataset of the cell-cycle-regulated gene prediction, we evaluated the effects on the accuracy of the benchmark gene sets prediction. The benchmark sets of the cell-cycle-regulated genes were selected according to the following criteria [4,7]: 1) the cell-cycle-regulated genes previously identified from small scale biological experiments; 2) the genes whose promoters were bound by at least one of the known cell cycle transcription factors in the Chromatin IP studies, excluding the genes in 1); and 3) genes annotated in MIPS as cell cycle and DNA processing, excluding the genes in 1) and 2). For budding yeast, there are 113 genes in the first benchmark set, 352 in the second, and

518 in the third. For fission yeast, there are 40 genes in the first benchmark set, 188 in the second, and 321 in the third. The accuracy of the predicted cell-cycle-regulated genes is the fraction of genes in the benchmark sets predicted in the 100~800 top scored genes.

4 Results

We compared the missing value estimation methods on the real application of cell-cycle-regulated gene prediction and the simulated missing data from the real microarray data.

4.1 Data

The data sets we used are the budding yeast data from Spellman et al [13] and the fission yeast data from Rustici et al [10].

4.1.1 Data from budding yeast

In Spellman et al's budding yeast data [13], two experiments are alpha data and cdc15 data and the cdc28 data was from Cho et al [2]. The properties of the data are summarized in Table 2. Table 2 shows that around 27%~77% genes have missing values.

Table 2 Summary of the budding yeast microarray data

	Alpha data	**Cdc15 data**	**Cdc28 data**
Number of genes	6178	6178	6178
Number of samples	18	24	17
Sampling interval (mintues)	7	10	10
Number of genes with missing values	1689 (27.3%)	1797 (29.1%)	4795 (77.6%)
Number of genes with complete values	4489 (72.7%)	4381 (70.9%)	1383 (22.4%)

Table 3 Summary of the fission yeast microarray data

	Cdc25-1	**Cdc25-21**	**Cdc25-22**	**Cdc25-sep1**	**Elu-1**	**Elu-2**	**Elu-3**
Number of genes	4991	4991	4991	4991	4991	4991	4991
Number of samples	19	18	18	20	20	20	20
Sampling interval (minutes)	15	15	15	15	15	15	15
Number of genes with missing values	1764 (35.3%)	1526 (30.6%)	1782 (35.7%)	2347 (47.0%)	1466 (29.4%)	1477 (29.6%)	2044 (41.0%)
Number of genes with complete values	3227 (64.7%)	3465 (69.4%)	3209 (64.3%)	2644 (53.0%)	3525 (70.6%)	3514 (70.4%)	2947 (59.0%)

4.1.2 Data from fission yeast

The fission yeast microarray data is from Rustici et al [10]. In their data, there are two technical repeats in cdc25-2 experiments. We treat them differently as cdc25-21 and cdc25-22. The properties of the data are summarized in Table 3. Table 3 shows that around 29%~47% genes in the fission yeast data have missing values.

4.1.3 Simulated missing data

For the budding yeast and fission yeast microarray data, we generated the simulated missing data from the genes with complete values. First, we filtered the original data to get only the genes with complete values. The numbers of the genes with complete data in

different experiments are summarized in Table 2 and Table 3. The percents of the simulated missing values are chosen as 1%, 5%, 10%, 15% and 20%. For each percentage, we generated 10 different missing data sets at random for validation. Therefore there will be 150 simulated missing data sets for budding yeast, and 350 simulated missing data sets for fission yeast.

4.2 Results of the improved knn-based method

The improved knn-based method with one iteration is similar as the naïve knn-based method. The experiments of the improved knn-based methods with different iterations are compared in the simulated missing data. The results of the simulated missing data from budding yeast's cdc15 data are shown in Figure 2. Figure 2 shows that the iterative process to fill in the missing values can reduce the RMSE and NRMSEs. Especially when there are significant percents of missing values in the data, the reduction in RMSE and NRMSEs are more obvious (around 10%). The possible explanation is that the nearest neighbors from the data with less missing values do not change much, and the method can find better nearest neighbors in the process when the dataset contains more missing values. Figure 2 also shows that, after 3 iterations, RMSE and NRMSEs converge to certain values.

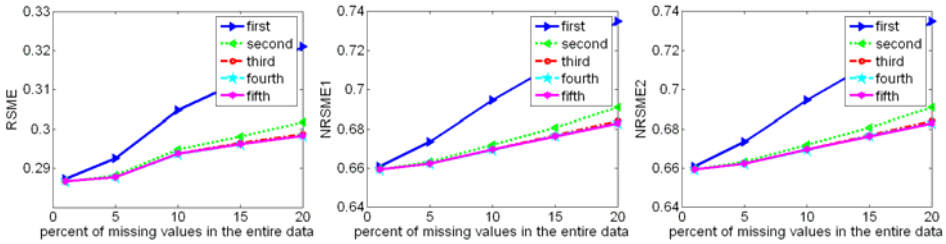


Figure 2. Comparison of the improved knn-based method with different iterations

4.3 Results of the different methods on the simulated missing data

The results of the different methods on the simulated missing data are shown in Figure 3 and Figure 4. For each original complete data and each missing percentage, the shown results are the mean of the 10 different simulated missing data. The results show that BPCA are always the bests in terms of RMSE, NRMSE1 and NRMSE2. LLSimpute achieves better results in fission yeast data when the percent of the missing data is small. The worse results are from SplineImpute for budding yeast data and from ROWimpute for fission yeast data. The performances of KNNimpute and IKNNimpute are in the middle, although IKNNimpute is a little bit better than KNNimpute.

The experiments also show that NRMSE1 is an appropriate measure to compare the performance of the methods on different data (the results are not shown). The reason is that, for the simulated missing data in the experiments, NRMSE1 from ROWimpute is around 1 for different data sets. This makes its results from different original data comparable. RSME from ROWimpute always changes for different data, and NRMSE2 is dependent on the original data – It is similar as NRMSE1 in the budding yeast data,

and similar as RMSE in the fission yeast data. This shows that NRMSE1 is well normalized and NRMSE2 is not.

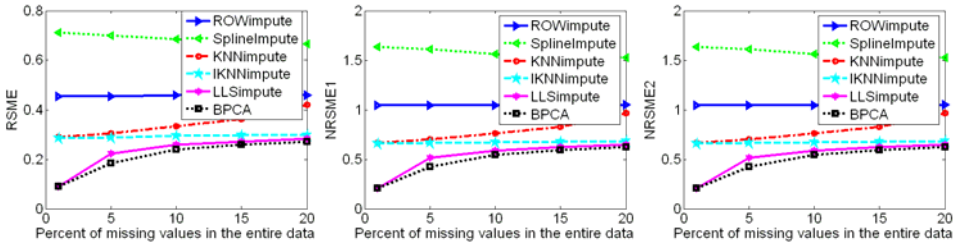


Figure 3. Performance of six methods on the simulated missing data from Spellman et al's data.

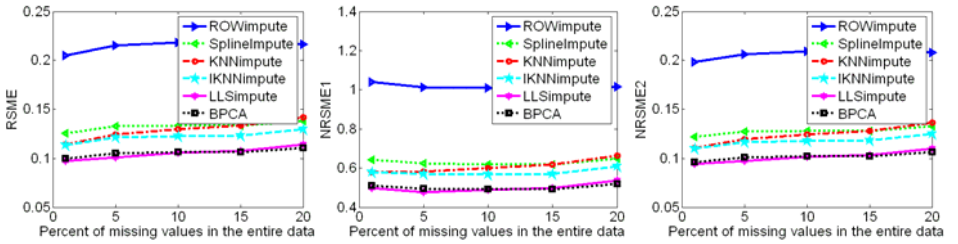


Figure 4. The performance of the six methods on the simulated missing data from Rustici et al's data.

4.4 Results of the different methods on the cell-cycle-regulated genes prediction

The prediction accuracies of the cell-cycle-regulated genes are shown in Figure 5 (budding yeast) and Figure 6 (fission yeast, see supplementary materials in http://www.comp.nus.edu.sg/~ligl/missing_values/). The x axis in Figure 5 and 6 represents the number of the genes with the top cdc scores predicted as cell-cycle-regulated genes. The y axis represents the fraction of the genes in the benchmark sets are predicted as cell-cycle-regulated genes.

To our surprise, the effects of the most methods on the cell-cycle-regulated gene prediction are similar, although some methods perform obviously better in terms of RMSE, NRMSE1 and NRMSE2. The simple row-mean method can achieve the similar accuracy as KNNimpute, IKNNimpute and BPCA. In both budding yeast data and fission yeast data, LLSimpute is the only exception. It performs quite well on the simulated missing data. However, the prediction accuracies of cell-cycle-regulated genes are quite low when the missing values are estimated with LLSimpute. The possible explanation is that LLSimpute overfits the data in the missing data estimation, since it always chooses over 1000 genes as the nearest neighbors.

5 Conclusion and discussions

Missing values are a common property in the microarray data. In this paper, we compared six different missing value estimation methods with microarray data from budding yeast and fission yeast. The results from the simulated missing data show that BPCA and LLSimpute can achieve much better RMSE, NRMSE1 and NRMSE2 than

other tested methods. And our improved KNN-based method can perform better than or at least as good as KNNimpute in terms of the three criteria.

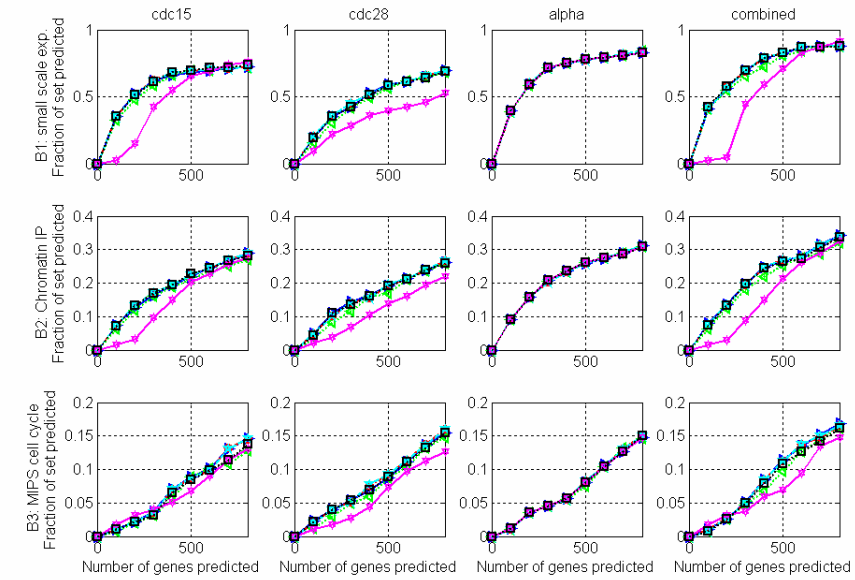


Figure 5. The performance of the six missing value estimation methods on the accuracy of the cell-cycle-regulated genes prediction from Spellman et al's budding yeast data. Legend: ROWimpute (blue solid line with right triangle), SplineImpute (green dotted line with left triangle), KNNimpute (red dashdot line with circle), IKNNimpute (cyan dashed line with pentagram), LLSimpute (magenta solid line with hexagram), BPCA (black dotted line with Square)

However, the results from the real application show that such improvement in the NRMSE does not benefit the cell-cycle-regulated gene prediction. Especially, the local least square method LLSimpute performs much worse on the cell-cycle-regulated gene prediction in our experiments, although it shows quite good NRMSE on the simulated missing data. And to our surprise, we found that the very basic method (ROWimpute) performs as good as other sophisticated methods on the real dataset of cell-cycle-regulated gene prediction. These results suggest that a new method for missing value estimation should be tested on both the simulated missing values and the real datasets for the validation purpose.

In literature, there are works to examine the effects of the missing value estimation methods on the real applications [3,11]. De Brevern et al [3] examined the influence of missing value estimation on the stability of the hierarchical clustering with microarray data. Their experiments show that the presence of the missing values has a significant effect on the gene cluster instability. Scheel et al [11] examined the influence of missing value estimation on the differentially expressed gene prediction. Their experiments show that the presence of the missing values leads to the loss of the differentially expressed genes predicted. Together with our results, their studies suggest that the smaller RMSE

and NRMSE are not the only criteria for missing value estimation and the missing value estimation methods should be used carefully in the real applications.

Reference:

[1] J. Chen, Identification of significant periodic genes in microarray gene expression data, BMC Bioinformatics 6 (2005) 286.

[2] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, R.W. Davis, A genome-wide transcriptional analysis of the mitotic cell cycle, Molecular Cell 2 (1998) 65-73.

[3] A.G. de Brevern, S. Hazout, A. Malpertuy, Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering, BMC Bioinformatics 5 (2004) 114.

[4] U. de Lichtenberg, L.J. Jensen, A. Fausboll, T.S. Jensen, P. Bork, S. Brunak, Comparison of computational methods for the identification of cell cycle-regulated genes, Bioinformatics 21 (2005) 1164-1171.

[5] X. Gan, A.W. Liew, H. Yan, Microarray missing data imputation based on a set theoretic framework and biological knowledge, Nucleic Acids Res 34 (2006) 1608-1619.

[6] H. Kim, G.H. Golub, H. Park, Missing value estimation for DNA microarray gene expression data: local least squares imputation, Bioinformatics 21 (2005) 187-198.

[7] S. Marguerat, T.S. Jensen, U. de Lichtenberg, B.T. Wilhelm, L.J. Jensen, J. Bahler, The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast, Yeast 23 (2006) 261-277.

[8] K.R. Murthy, J.H. Liu, Improved fourier transform method for unsupervised cell-cycle regulated gene prediction, in: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (2004) 194-203.

[9] S. Oba, M.A. Sato, I. Takemasa, M. Monden, K. Matsubara, S. Ishii, A Bayesian missing value estimation method for gene expression profile data, Bioinformatics 19 (2003) 2088-2096.

[10] G. Rustici, J. Mata, K. Kivinen, P. Lio, C.J. Penkett, G. Burns, J. Hayles, A. Brazma, P. Nurse, J. Bahler, Periodic gene expression program of the fission yeast cell cycle, Nat Genet 36 (2004) 809-817.

[11] I. Scheel, M. Aldrin, I.K. Glad, R. Sorum, H. Lyng, A. Frigessi, The influence of missing value imputation on detection of differentially expressed genes from microarray data, Bioinformatics 21 (2005) 4272-4279.

[12] M.S. Sehgal, I. Gondal, L.S. Dooley, Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data, Bioinformatics 21 (2005) 2417-2423.

[13] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, Molecular Biology of the Cell 9 (1998) 3273-3297.

[14] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics 17 (2001) 520-525.

[15] X. Wang, A. Li, Z. Jiang, H. Feng, Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme, BMC Bioinformatics 7 (2006) 32.

[16] X. Zhou, X. Wang, E.R. Dougherty, Missing-value estimation using linear and non-linear regression with Bayesian gene selection, Bioinformatics 19 (2003) 2302-2307.