[Frontiers in Bioscience 12, 1845-1851, January 1, 2007]

Identification of differentially expressed genes in multiple microarray experiments using discrete fourier transform

Keng Wah Choo, Waiming Kong

Bioinformatics Group, Nanyang Polytechnic, 569830 Singapore

TABLE OF CONTENTS

1. ABSTRACT

Research in the post-genome sequence era has been shifting towards a functional understanding of the roles and relationships between different genes in different conditions. While the advances in genetic expression profiling techniques including microarrays enable detailed and genome-scale measurements, the extraction of meaningful information from large datasets remains a challenging task. Here, we propose a novel method of generating gene differential expression profiles such that gene expression values from one dataset can be directly compared with those of another dataset. A simplified Discrete Fourier Transform is applied to interposed gene expression values, thereby generating the 'spectra' for a pair of conditions. Using this technique, differentially expressed genes produce higher amplitudes at the Nyquist Frequency. By measuring the phase of the 'spectra' generated, the over- and under-expressed nature of the genes can be identified. This method was validated using two sets of GeneChip array data, one from prostate cancer related dataset and the other from macular degeneration related dataset. The genes identified as differentially expressed by our method were found to be similar to those published using their preferred methods. Based on our findings, the proposed DFT method could be used efficiently in identifying differentially expressed genes from multiple-array experiments from two different conditions.

2. INTRODUCTION

The application of Discrete Fourier Transform (DFT) Theory in the analysis of microarray gene-expression data has been limited thus far. The theory has been applied in model-based methods of Luan and Li for identifying temporally expressed genes based on time course microarrays (1); DFT method of Wichert *et al.* in identifying temporally expressed transcripts in microarray time series data (2); Fourier harmonic approach of Zhang *et al.* for visualizing temporal pattern of gene expression (3); and DFT method of Shedden and Cooper for analysis of cell-cycle-specific gene expression in human cells (4). It has become apparent that the cell-cycle specific genes are expressed in a temporal manner so that the Fourier Transform Theory can be applied. The limited DFT applications may be due to the lack of naturally formed waveforms within microarray expression data.

Depending on the underlying biological conditions, only a handful of genes are expected to have significant changes in their expression levels during the course of a microarray experiment. Many approaches have been described in the literature to detect the differentially expressed genes. These include fold change test and t-test by Long *et al.*(5), significance analysis of microarray (SAM) by Tusher *et al.*(6) and ANOVA by Kerr *et al.*(7). The fold change test has been popular mainly due to its simplicity. In this method, the potential candidates of differentially expressed genes are selected based on large fold change variation. If a selection is based on a single slide, the confidence level is low because real changes cannot be distinguished from experimental artifacts. Thus, the selected candidates of differentially expressed genes from single experiment have to be verified by repeating the experiment or by conducting experiments employing other methods such as Northern blotting.

Various parametric-based models and nonparametric-based models were also developed to address the problem. In 2002, Pan W *et al.* investigated a model-based cluster analysis of microarray gene-expression data (8). Pan W also described a new approach in 2004, using permutation and nonparametric methods to analyze differential gene expression (9). Jeffrey *et al.,* on the other hand, proposed a statistical modeling approach to uncover differentially expressed genes (10).

Although each method has its own merits and produces reasonably good results, a good grasp of the complex models and knowledge in statistics is necessary to ensure correct and effective use of these approaches. Hence, a search for a simple and yet effective method is important for the advancement in this field of research. The method proposed here aims to achieve this very objective by applying a simplified DFT on intuitively arranged gene expression intensities from multiple microarray experiments.

3. MATERIALS AND METHODS

The proposed method uses Fourier Transform (FT) Theory and Sampling Theorem in detecting differentially expressed genes from microarray data. Since the data is processed in a computer, the Discrete FT is used instead of continuous FT. With a unique arrangement of the gene expression intensities and the use of Parseval's Theorem, the original DFT equation can be further simplified for ease of implementation and fast computation.

3.1. Discrete Fourier Transform

The Sampling Theorem states that for a limited bandwidth signal with maximum frequency component f_{max}, the sampling frequency (f_s) used for capturing the signal must be greater than twice of that f_{max} (i.e., $f_s > 2f_{max}$), to prevent aliasing. The frequency $2f_{max}$ is called the Nyquist sampling rate and half of this value, which is f_{max}, is referred to as the Nyquist frequency. The Sampling Theorem was introduced by Nyquist in 1928 and mathematically proven by Shannon in 1949. The terms 'Nyquist Sampling Theorem' and 'Shannon Sampling Theorem' are used interchangeably. They are in fact the same Sampling Theorem. The Fourier transform, in essence, decomposes or separates a waveform or function into sinusoids of different frequencies which sum to the original waveform. It identifies or distinguishes the different sinusoids and their respective amplitudes. The Fourier transform of a signal *f(x)* is defined as:

$$F(s) = \int_{-\infty}^{\infty} f(x) \exp(-i\, 2\pi x s)\, dx. \qquad (1)$$

Since a digital computer works only with discrete data, numerical computation of the Fourier transform of *f(x)* requires discrete sample values of *f(x)*, which is called f_k. In addition, a computer can only compute the Fourier transform *F(s)* at discrete values of *s*, that is, it can only provide discrete frequency samples of the transform, F_r. If $f(kT)$ and $F(rs_0)$ are the k^th^ and r^th^ samples of *f(x)* and *F(s)* respectively, and N_0 is the number of samples in the signal in one period T_0, then f_k is defined as:

$$f_k = T f(kT) = T_0 N_0^{-1} f(kT) \qquad (2)$$

$$F_r = F(rs_0) \qquad \text{and}$$

$$s_0 = 2\pi \mathcal{F}_0 = 2\pi T_0^{-1}.$$

The Discrete Fourier Transform (DFT) of *f(x)* is therefore given by:

$$F_r = \sum_{k=0}^{N_0-1} f_k \exp(-i\, r\Omega_0 k) \qquad (3)$$

where $\Omega_0 = 2\pi N_0^{-1}$.

3.2. Arrangement of Gene Expression Intensity and Its Transform

The proposed method describes a simple way of arranging two sets of microarray gene-expression data. The two sets of data could be derived from tissues under a pair of conditions, such as tumor versus normal or tumor *x* versus tumor *y*. For simplicity, the letters *T* and *N* are used to represent the two conditions. If there are *M* arrays, the gene-expression intensities are arranged in an interposed order, i.e., $(T_1N_1)(T_2N_2)\ldots(T_{M-1}N_{M-1})(T_MN_M)$ such that those of the second condition interpose the expression intensities from first condition. Table 1 shows a possible arrangement for *M* arrays of *N* genes each.

If a gene is consistently over-expressed in the *T* slides as compared to the *N* slides, the values in the columns T_i will be constantly higher than those in the columns N_i. The alternating expression intensities in *T* and *N* slides will then produce a consecutive high-low profile, which entails high DFT magnitude at the Nyquist frequency component. In Figure 1, Data A is a plot of intensities of differentially expressed gene and Data B that of a non-differentially expressed gene. The magnitudes of the DFT on these two datasets are shown in Figure 2. The frequency with the highest magnitude is f_{max} located at the $(M/2)^{th}$ position. This frequency component is the Nyquist frequency. Likewise, if a gene is consistently under-expressed in *T* slides *vis-à-vis* the *N* slides, the alternating *T* and *N* slides will produce a consecutive low-high profile, which will generate high DFT magnitude at the Nyquist frequency component. In other words, a gene that is differentially expressed will present a zigzag waveform producing high DFT magnitude at the Nyquist frequency.

A quantitative measure of differentially expressed genes under two different conditions is described here. We normalize the magnitude of the Nyquist Frequency component to the sum of all the components. The normalized amplitude is defined as:

Table 1. Interposed arrangement of expression intensity for *M* arrays and *N* genes under *T* and *N* conditions

Genes	**T_1**	**N_1**	**T_2**	**N_2**	**T_{M-1}**	**N_{M-1}**	**T_M**	**N_M**
Gene1	393.15	280.4	323.75	316.9	327	300.5	255.15	263.9
Gene2	151.4	102.5	108.5	117	112	88.4	108.65	103
Gene3	137.65	103.65	101.75	107.9	113	86.15	102.5	91.15
Gene4	217.65	189.75	213.5	179.65	223.15	199.05	177.4	170.4
Gene5	191.4	135	152.5	143.9	150.65	146.15	146.15	134
Gene6	929.25	122	141.25	197.15	143.75	152.15	171.25	129.5
Gene n-1	141.15	98	98.65	95.15	112.15	82.4	118.75	100.15
Gene N	582.65	210.75	259	387.25	234	285.25	360.65	215.75

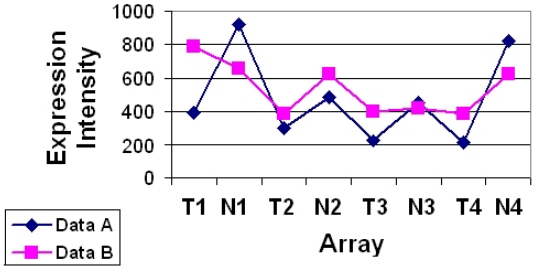


Figure 1. Gene expression intensities from 4 pair of slides arranged in *T-N* manner: Data A, representing a differentially expressed gene, has a distinctive zigzag profile; Data B does not have such profile.

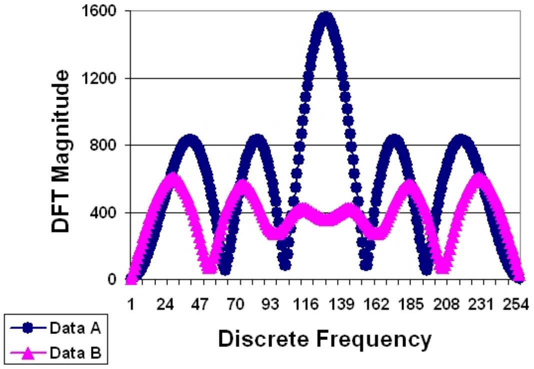


Figure 2. Plot of the magnitude of 256-point DFT performed on Data A (DFT A) and Data B (DFT B): DFT of Data A with zigzag profile has higher value at the Nyquist frequency as compared to that of Data B.

$$F_{r_N} = \frac{|F_r|}{\sqrt{F_{Total}}} \tag{4}$$

where

$$F_{Total} = \sum_{r=0}^{N_0-1} |F_r|^2 \tag{5}$$

Making use of the Parsevals's Theorem which states that:

$$\sum_{r=0}^{N_0-1} |F_r|^2 = N_0 \sum_{k=0}^{N_0-1} |f_k|^2 \tag{6}$$

The normalized amplitude becomes:

$$F_{r_N} = \frac{|F_r|}{\sqrt{N_0 \sum_{k=0}^{N_0-1} |f_k|^2}} \tag{7}$$

This simplification leads to a significant reduction in computational time. Instead of computing the complete N_0 points of DFT, we only need to compute one point, that of the Nyquist frequency component, and normalize it with the sum of the gene-expression as shown in equation (7). This normalized amplitude gives the ratio of the Nyquist Frequency magnitude to the sum of magnitudes. Indirectly it gives a score on the consistency of a gene being differentially expressed over the *M* slides in the microarray experiments.

In addition, a DFT on any data series results in complex values, i.e. $F_k = R_k + J_k i$, where R_k is the real value and J_k is the imaginary value of the transformed F_k, *i* defined as √(-1), represents the imaginary term. The phase of the component F_k is computed as:

$$\theta_k = \tan^{-1}(\frac{J_k}{R_k}) \tag{8}$$

Figure 3 shows two expression datasets, Data A represents under-expressed gene in *T* as compared to *N* condition, and Data C represents over-expressed gene in *T-N* conditions. Figure 4 shows the plot of the phase for the two transformed data series. It is interesting to note that the phase for the Nyquist component is -*π* radians for Data A (the under-expressed gene) and *zero* radian for Data C (the over-expressed gene). This property can be exploited in that a phase of *zero* radian at Nyquist component indicates an over-expressed gene, and that a phase of -*π* radian indicates an under-expressed gene. This feature removes the need to compute the ratios or the log ratios used in the popular fold-change method. The phase and normalized amplitude of the Nyquist Frequency component are sufficient for ferreting differentially expressed genes and determining whether they are over- or under-expressed genes.

3.3. The Scatter Plot of Normalized Amplitude Versus Absolute Amplitude

For the purpose of comparison, three common representations of differential gene expression are shown in Figure 5. Figure 5a shows a scatter plot of microarray data where the Red intensity (R) is plotted against the Green intensity (G), the R and G intensities represent gene

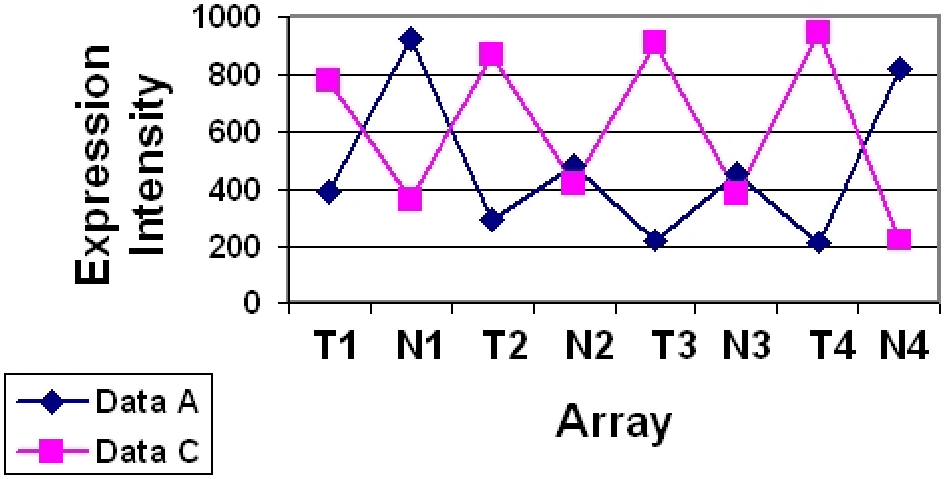


Figure 3. Gene expression intensities from 4 pair of slides arranged in *T-N* manner: Data A, representing under-expressed gene in *T-N* condition, has a low-high profile; Data C, representing over-expressed gene, has a high-low profile.

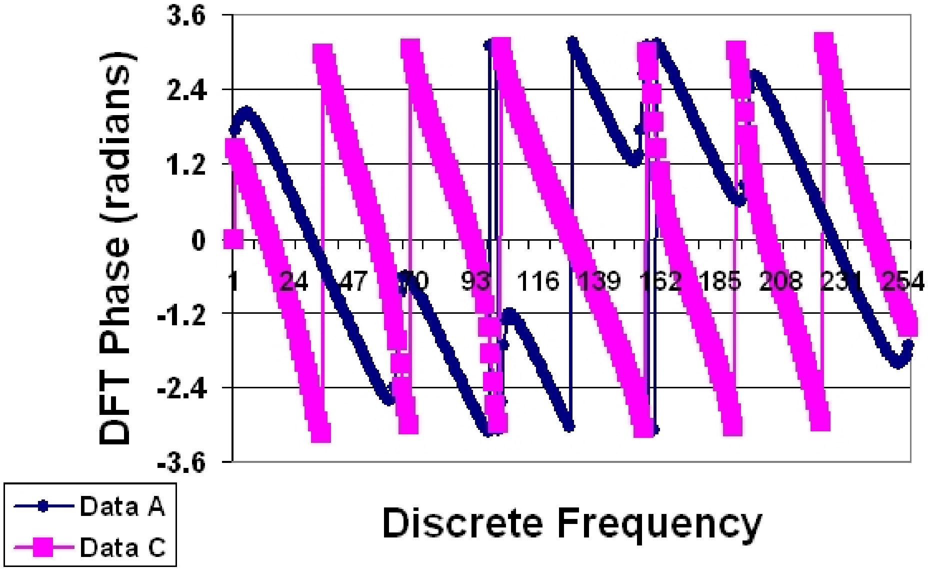


Figure 4. Phase plot of 256-point DFT performed on data shown in Figure 3 after removing the mean value from these data points. The phase of Nyquist Frequency component located at 129 is -π radians for Data A and zero radian for Data C. A measure on the phase of one frequency component reveals whether a gene is over-expressed or under-expressed in the datasets.

expressions obtained from diseased and normal conditions respectively. Figure 5b shows another scatter plot of microarray data where the logarithm of Red intensity ($\log_2 R$) is plotted against the logarithm of Green intensity ($\log_2 G$). Figure 5c shows a third scatter plot of microarray data where M is plotted against A. M is defined as $\log_2 C$, where C = R/G, and A as $(\log_2 R + \log_2 G)/2$.

The use of DFT magnitude and phase presents more information visually. Figure 6 shows a plot where the normalized amplitude of the Nyquist frequency is plotted against the absolute amplitude. The normalized amplitudes of over-expressed genes, with phase equals *zero* radian , are set to positive values, and the normalized amplitudes of under-expressed genes, with phase equals $-\pi$ radian, are set to negative values. In this novel representation, the differentially expressed genes are prominent and readily identified. In addition, the higher the magnitude of the normalized frequency component, the more consistent is the gene expression profile over the *M* slides. Similarly, the

Table 2. Prediction result from (1)

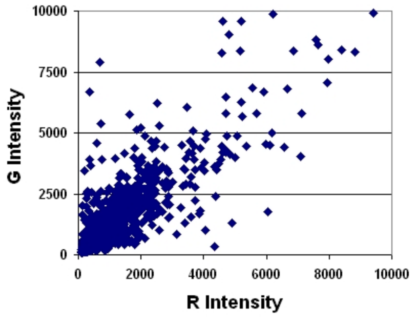
	Raw data				Normalized data			
4 Gene model		TT	TN	85.7%		TT	TN	77.2%
	PT	26	4	(30/35)	PT	19	0	(27/35)
	PN	1	4	p=0.0006[1]	PN	8	8	p=0.0005
16 Gene Model		TT	TN	82.9%		TT	TN	85.7%
	PT	27	6	(29/35)	PT	22	0	(30/35)
	PN	0	2	p=0.047	PN	5	8	p=0.0001

TT = True Tumour, TN = True Normal, PT = Predicted Tumour, PN = Predicted Normal, [1] All values calculated by Fisher's Exact Test

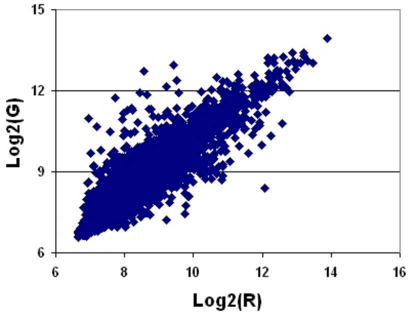
Table 3. Prediction result using the DFT method

	Raw data			Normalized data	
	Disease	Normal		Disease	Normal
True Test	39	2	**True Test**	42	2
False Test	7	44	False Test	4	44

(a) Plot R Vs G



(b) Plot Log2(R) Vs Log2(G)



(c) Plot M Vs A

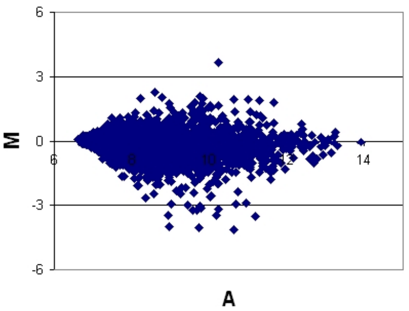


Figure 5. Three different gene-expression plots visualizing the ratio and the $\log_2$(ratio) of gene expression intensities.

higher the magnitude of absolute amplitude, the higher is the fold changes between the T and N conditions. The novel plot also presents different quadrants that represent gene expression of different characteristics; genes located in the high normalized-amplitude and high absolute-amplitude quadrant will be best for use in clustering the microarray data. This plot complements well with the plots mentioned earlier.

4. RESULTS

To evaluate our method, we apply it to two published microarray datasets – a prostate cancer related dataset and a macular degeneration related dataset.

4.1. Performance on Prostate Cancer Datasets

The first dataset, the prostate cancer related dataset, was obtained from URL: http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=75. In this study (11), microarray expression analysis was used to identify genes that might anticipate the clinical behavior of prostate cancer. Singh et al. identified a set of genes that strongly correlated with the state of tumor differentiation, measured by Gleason score. The prediction result is shown in Table 2.

This dataset contains gene expression data from 46 Affymetrix slides of prostate cancer tumor tissues and 46 slides of normal tissues. The expression intensities of the prostate cancer tissue and normal tissue were interposed in the manner as shown in Figure 1. These interposed intensities were then subjected to DFT as described in the previous section. We identified the top 10 genes and used them for classification with Eisen's K-Means Cluster Program (12). We successfully isolated the tumor slides from the normal slides with high prediction rate, as shown in Table 3. Table 3a shows the prediction result without normalizing the data (raw values) and Table 3b shows the prediction result on mean-normalized data (i.e. the data is mean-normalized before using the proposed method). The prediction results, based on true-positive and true-negative tests, are 90.21% and 93.48% respectively for raw and normalized data. It is worth noting that 9 of our 18 top ranking genes are identical to those in the 16-gene model used by Singh et al. Table 4 summarizes the top

Table 4. Top 18 ranking genes identified by DFT method as compared to Singh's 16-gene prediction model

S/N	Affy ID	Singh's Model	S/N	Affy ID	Singh's Model
1	37639_at	x	10	39315_at	
2	38634_at	x	11	40282_s_at	x
3	41152_f_at		12	1804_at	
4	32598_at	x	13	31527_at	
5	41706_at		14	38406_f_at	x
6	37366_at		15	39755_at	x
7	33656_at		16	40024_at	
8	41468_at	x	17	39054_at	x
9	1740_g_at		18	38028_at	x

Affy ID = Affymetrix Probe ID, x = found in Singh's 16-gene prediction model (11).

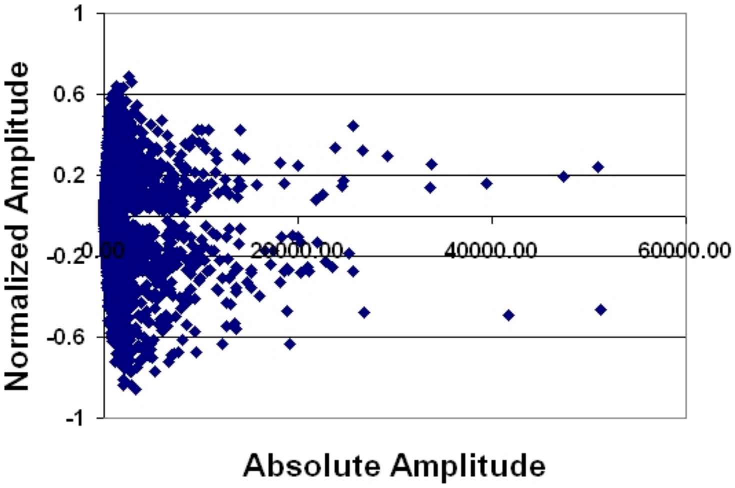


Figure 6. Scatter plot of the Nyquist Frequency normalized amplitude and absolute amplitude.

18 genes identified using our method compared to those in Singh's model.

4.2. Performance on Age-Related Macular Degeneration Datasets

The second gene-expression data, macular disease related dataset, was obtained from public database published at the following URL: http://microarray.cnmcresearch.org/ListProjExp.asp?ProjectName=WSilk+Macular+Degeneration, by Dr Karl GC. He conducted this experiment entitled 'Age-related macular degeneration has a strong epidemiological association with cardiovascular disease' at the National Eye Institute with the aim to identify disease-specific genes. Using microarray technology, 13 gene-expression data each were obtained from diseased and age-matched control patients. We applied our method to these data and identified top 10 ranking genes using our scoring method. These ten genes achieved 100% prediction rate in differentiating diseased samples from the control ones.

5. DISCUSSION

We discussed and presented two hypotheses. The first hypothesis describes a simplified DFT method that can be used for fast and efficient detection of differentially expressed genes from multiple microarray experiments. The second hypothesis reveals a new way in visualizing multiple microarray gene-expression intensities by plotting the normalized amplitude, incorporating phase information, against the absolute amplitude of the transformed intensities. It was shown in two real microarray datasets that this approach was able to detect consistently over- and under-expressed genes. It also showed that the method worked on both normalized and raw datasets. The scatter plot allows us to visualize genes of different expression modes plotted in different quadrants. Finally, the approach assumes that equal numbers of microarray experiments for tissues under two different conditions are conducted.

6. ACKNOWLEDGEMENTS

The author would like to thank the institutions for allowing public access to their microarray data. The author is grateful to Nanyang Polytechnic for the continuous support.

7. REFERENCES

1. Luan Y. and H. Li: Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics* Vol 20, 332–339 (2004)

2. Wichert S, K. Fokianos and K. Strimmer: Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics* Vol. 20, 5-20 (2004)

3. Zhang L, A. Zhang and M. Ramanathan: Fourier harmonic approach for visualizing temporal patterns of gene expression data. *The Computational Systems Bioinformatics Conference* (2003)

4. Shedden K. and S. Cooper: Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc Natl Acad Sci* U S A. 99(7), 4379–4384 (2002)

5. Long A.D, H.J. Mangalam, B.Y.P. Chan, L. Tolleri, G.W. Hatfield and P. Baldi: Improved statistical inference from dna microarray data using analysis of variance and a bayesian statistical framework. *J Biol Chem* Vol. 276, Issue 23, 19937-19944 (2001)

6. Tusher V.G, R. Tibshirani and G. Chu: Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* 98: 5116–5121 (2001)

7. Kerr MK, M. Martin M, G.A. Churchill: Analysis of variance for gene expression microarray data. *J Comp Biol* 7: 819-837 (2000)

8. Pan W, J. Lin and C. Le: Model-based cluster analysis of microarray gene-expression data. *Genome Biology,* 3(2): research0009.1-0009.8 (2002)

9. Pan W: On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics* Vol. 19, 1333-1340 (2004)

10. Jeffrey G.T, J.M. Olson, S.J. Tapscott and L.P. Zhao: An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res* 11: 1227-1236 (2001)

11. Singh D, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers: Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* Vol. 1, 203-209 (2002)

12. Eisen M.B, P.T. Spellman, P.O. Brown, and D. Botstein: Cluster analysis and display of genome-wide expression patterns. *Genetics* Vol. 95, Issue 25, 14863-14868 (1998)

Key Words: Discrete Fourier Transform, Microarray Data, and Differentially Expressed Genes

Send correspondence to: Dr Keng Wah Choo, Bioinformatics Group, Nanyang Polytechnic, 569830 Singapore, Tel: 65-65500-587, Fax: 65-64520-400, E-mail: CHOO_Keng_Wah@nyp.gov.sg

http://www.bioscience.org/current/vol12.htm