

---

# Spectral Analysis of Microarray Gene Expression Time Series Data of *Plasmodium Falciparum*

---

Liping Du<sup>1,2</sup>, Shuanhu Wu<sup>1,3</sup>, Alan Wee-Chung Liew<sup>4</sup>, David K. Smith<sup>5</sup>, and Hong Yan<sup>1,6</sup>

1. Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

2. School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China

3. School of Computer Science and Technology, Yantai University, Shandong, Yantai 264005, China

4. School of Information and Communication Technology, Griffith University, Gold Coast Campus, QLD4222, Queensland, Australia

5. Department of Biochemistry, University of Hong Kong, Pok Fu Lam, Hong Kong

6. School of Electronic and Information Engineering, University of Sydney, NSW2006, Sydney, Australia

E-mail: [dlp2001@ies.ustb.edu.cn](mailto:dlp2001@ies.ustb.edu.cn), [wushuanhu@gmail.com](mailto:wushuanhu@gmail.com), [a.liew@griffith.edu.au](mailto:a.liew@griffith.edu.au) (corresponding author), [dsmith@hku.hk](mailto:dsmith@hku.hk), [h.yan@cityu.edu.hk](mailto:h.yan@cityu.edu.hk)

**Abstract:** We propose a new strategy to analyze the periodicity of gene expression profiles using singular spectrum analysis (SSA) and autoregressive (AR) model based spectral estimation. By combining the advantages of SSA and AR modeling, more periodic genes are extracted in the *Plasmodium falciparum* dataset compared with the classical Fourier analysis technique. We are able to identify more gene targets for new drug discovery, and by checking against the seven well known malaria vaccine candidates, we have found 5 additional genes that warrant further biological verification.

**Keywords:** Singular spectrum analysis (SSA), autoregressive (AR) model, microarray time series analysis, gene target, *Plasmodium falciparum*

**Bibliographical notes:** Liping Du is a lecturer at the University of Science and Technology Beijing. She received her PhD from the Beijing Institute of Technology in 2005. From 2005 to 2006 she was a senior research assistant at City University of Hong Kong. Her research interests are in the areas of spectral analysis and bioinformatics.

*L.P. Du, S.H. Wu, A.W.C. Liew, D.K. Smith, H. Yan*

Shuanhu Wu is a professor at the School of Computer Science & Technology of Yantai University, Yantai, Shandong, China. He received his Ph.D degree in Dec.2001 with a major of image processing and image coding in electronic and information engineering from Xi'an Jiaotong University, Xi'an, China. He joined Yantai University in 2003. During 2005 and 2006, he was a Research Fellow in the Department of Electronic Engineering, City University of Hong Kong. His current research interests include: genomic signal processing, signal and image processing and pattern recognition.

Alan Wee-Chung Liew received his B.Eng. with first class honors in Electrical and Electronic Engineering from the University of Auckland, New Zealand, in 1993 and Ph.D. in Electronic Engineering from the University of Tasmania, Australia, in 1997. From 1997 to 2004, he worked as a Research Fellow and later Senior Research Fellow at City University of Hong Kong. He was an Assistant Professor in the Department of Computer Science and Engineering, The Chinese University of Hong Kong from 2004-2007. Currently he is a Senior Lecturer in the School of Information and Communication Technology, Griffith University, Australia. His current research interests include computer vision, medical imaging, pattern recognition and bioinformatics. Dr. Liew is a senior member of the Institute of Electrical and Electronic Engineers (IEEE), and his biography is listed in Marquis Who's Who in the World and Marquis Who's Who in Science & Engineering.

David K. Smith obtained a BSc in Mathematics from the University of Queensland in 1976, an MA in Information Science from the University of Canberra in 1986 and a PhD in Biochemistry from the University of Melbourne in 1996. He designed and implemented computer systems for the Australian Government and managed bioinformatics core facilities in Universities before commencing a research career. Currently he is an Assistant Professor in the Department of Biochemistry at the University of Hong Kong. His research interests include applying computational methods to structural biology, gene expression studies and the role of intrinsically disordered proteins in signalling and tissue specificity.

Hong Yan received a B.E. degree from Nanking Institute of Posts and Telecommunications in 1982, an M.S.E. degree from the University of Michigan in 1984, and a Ph.D. degree from Yale University in 1989, all in electrical engineering. In 1982 and 1983 he worked on signal detection and estimation as a graduate student and research assistant at Tsinghua University. From 1986 to 1989 he was a research scientist at General Network Corporation, New Haven, CT, USA, where he worked on design and optimization of computer and telecommunications networks. He joined the University of Sydney in 1989 and became Professor of Imaging Science in 1997. He is currently Professor of Computer Engineering at City University of Hong Kong. His research interests include image processing, pattern recognition and bioinformatics. He is author, co-author or editor of two books and over 300 refereed technical papers in these areas. Professor Yan is a fellow of the Institute of Electrical and Electronic Engineers (IEEE), the International Association for Pattern Recognition (IAPR), and the Institution of Engineers, Australia (IEAust).

---

## **1. Introduction**

The protozoan parasite *Plasmodium falciparum* is one of the four *Plasmodium* species that can cause human malaria, for which there is no effective vaccine yet. The disease is a major killer in many developing countries. In 2002, a complete genome sequence of *P. falciparum* was reported (Gardner *et al.*, 2002) [3]. The database provides valuable information for researchers to find potentially unique or at least substantially different genes in *P. falciparum* compared with other species. These genes may be useful in designing drugs that can cause less risk of negative side effects.

Certain genes are expressed only at specific stages of the cell cycle, for example, the intraerythrocytic developmental cycle (IDC). These genes consequently exhibit a periodic pattern of expression. The identification of periodically expressed genes is important for understanding the biochemical functions and gene regulations of *P. falciparum*. Bozdech *et al.* have applied the Fourier transform to analyze the periodicity of transcriptome of IDC and offered a detailed description of the four major morphological stages, namely, ring/early trophozoite, trophozoite/early schizont, schizont and early ring (Bozdech *et al.*, 2003) [1]. Their paper revealed that a majority of expression profiles of transcriptome of the IDC showed a high periodicity. The spectral information can be used for further data analysis, such as clustering (Bozdech *et al.*, 2003; Spellman *et al.*, 1998; Le Roch *et al.*, 2003; Rajarajeswari *et al.*, 2005) [1,13,6,11]. Liew *et al.* have studied the problem of missing data and uneven sampling for the dataset (Liew *et al.*, 2007) [7]. However, the Fourier transform is well known to have limited frequency resolution for short signal analysis due to the so-called windowing or data truncation effect. As discussed below, there are only 46 time points in the gene expression profiles in the IDC data of *P. falciparum*. This kind of time series would be considered extremely short in signal processing. Furthermore, microarray data usually contain a high level of noise, which can degrade the performance of data analysis algorithms. Thus, effective methods are needed to process noisy and short gene expression time series data.

In this paper, we propose a new scheme for analyzing the periodicity of the transcriptome of the IDC by combining singular spectrum analysis (SSA) and autoregressive (AR) modeling. By using the SSA, the dominant trend can be extracted from the noisy expression profiles and the effect of noise can be reduced effectively. The method described here is an extension of our work presented at CompLife'06 (Du *et al.*, 2006) [2]. Utilizing the advantage of AR modeling in spectral analysis, we are able to analyze the periodicities of the data more accurately. The combination of SSA and AR methods can identify about 90% of genes in *P. falciparum* and thus gain an advantage over the Fourier analysis in cyclic gene identification.

## **2 Methods**

### *2.1 Dataset and data preprocessing*

The microarray dataset used in this paper is downloaded from <http://malaria.ucsf.edu>. It contains the expression profiles of 5080 oligonucleotides measured at 46 time points spanning 48 hours during the IDC with one hour time resolution for the HB3 strain. A logarithm transform was applied to the expression ratio of channel Cy5 to channel Cy3.

The mean of each expression profile was subtracted from each profile so that the average log ratio value over the time span is equal to zero.

## 2.2 Trend estimation of expression profiles using the SSA

Gene expression data generally contain high level of noise. Although spectral analysis can be applied directly to the original data, noise would degrade the results. A preprocessing step should be applied to the original data to reduce the effect of noise before performing spectral analysis.

Singular spectrum analysis (SSA) has been proven to be a powerful tool for processing many types of time series in geophysics, economics, biology, medicine and other sciences (Golyandina *et al.*, 2001) [4], which may have nonlinear characteristics. The main idea of SSA is to extract the underlying trend from short and noisy time series (Vautard *et al.*, 1992) [15]. There is no need to fit an assumed model to the time series since SSA is a robust model-free technique. These properties make SSA a useful technique for gene expression data processing. Using SSA, microarray dataset containing the expression profiles of 5080 oligonucleotides is processed to extract their trend curves and remove noise.

SSA performs singular value decomposition (SVD) on the so-called trajectory matrix obtained from the original time series with subsequent reconstruction of the series. The singular values can be grouped into two separate components: trend component and noise component. With the proper selection of singular values, the trend curve that represents the dominant spectral component can be reconstructed from the original expression profile. In fact, the process of SSA can be considered as a process of data fitting.

Let each expression profiles be a time series  $\{s_1, s_2, \dots, s_n, \dots, s_N\}$ . The SSA can be performed as follows:

(1) Construct the trajectory matrix  $X_{M,K}$  from the original series by sliding a window of length  $M$  ( $M \leq N/2$ ),  $K = N - M + 1$ .

$$X_{M,K} = (x_{ij} = s_{i+j-1}) = \begin{bmatrix} s_1 & s_2 & \cdots & s_K \\ s_2 & s_3 & \cdots & s_{K+1} \\ \vdots & \vdots & \cdots & \vdots \\ s_M & s_{M+1} & \cdots & s_N \end{bmatrix}$$

(2) Perform the SVD of the matrix  $R = XX^T$ . The singular values are ranked in decreasing order  $\lambda_i, (1 < i < M)$  ( $\lambda_1 > \lambda_2 > \cdots > \lambda_M$ ) and the trajectory matrix is decomposed into a series of components  $X_i$ , where  $X_i = \sqrt{\lambda_i} U_i V_i^T$  ( $i = 1, 2, \dots, M$ ) are rank-one biorthogonal matrices, the  $U_i$  and  $V_i$  are the left and right singular vectors of the matrix  $X$ , respectively.

(3) Group a specified number of leading singular values  $\lambda_i$  and sum the corresponding components  $X_i$ , then the resultant matrix is  $X'_{M,K} = (x'_{ij})$ . The number of singular values to use is discussed below.

(4) Reconstruct the data series  $\{s'_1, s'_2, \dots, s'_n, \dots, s'_N\}$  by averaging the elements of matrix  $X'$  over the 'diagonals'  $i + j = n + 1$ . The choice  $n = 1$  gives  $s'_1 = x'_{11}$ , for  $n=2$  we have  $s'_2 = (x'_{12} + x'_{21})/2$  and so on.

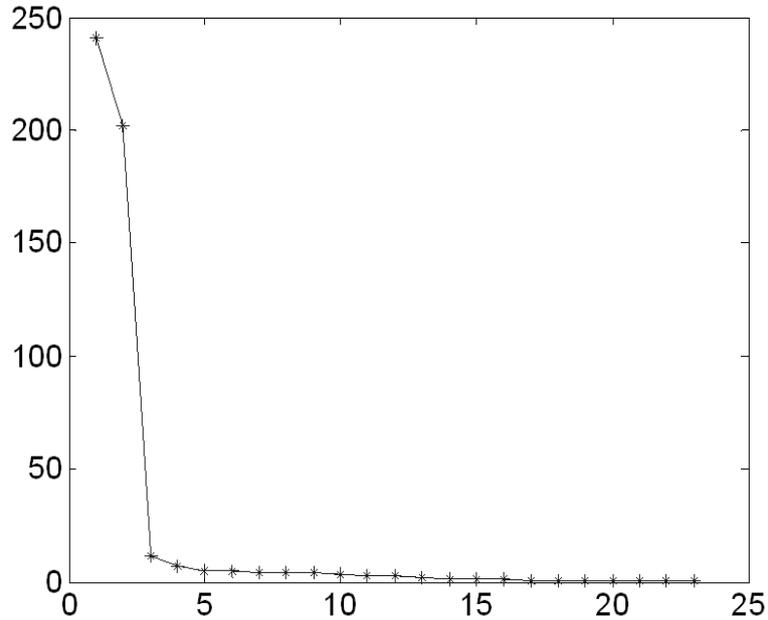
### *Spectral Analysis of Microarray Gene Expression Time Series Data*

One important consideration in SSA is how to select the singular values to reconstruct the expression profiles. If we plot the singular values, the graph contains an initial steep slope, representing the signal, and a “flat floor”, representing the noise level (Vautard *et al.*, 1989) [14]. The dominant component trend can be reconstructed from the expression profiles using the leading singular values that contain most of the energy. The ratio of the leading singular values and the total eigenvalues is defined as follows:

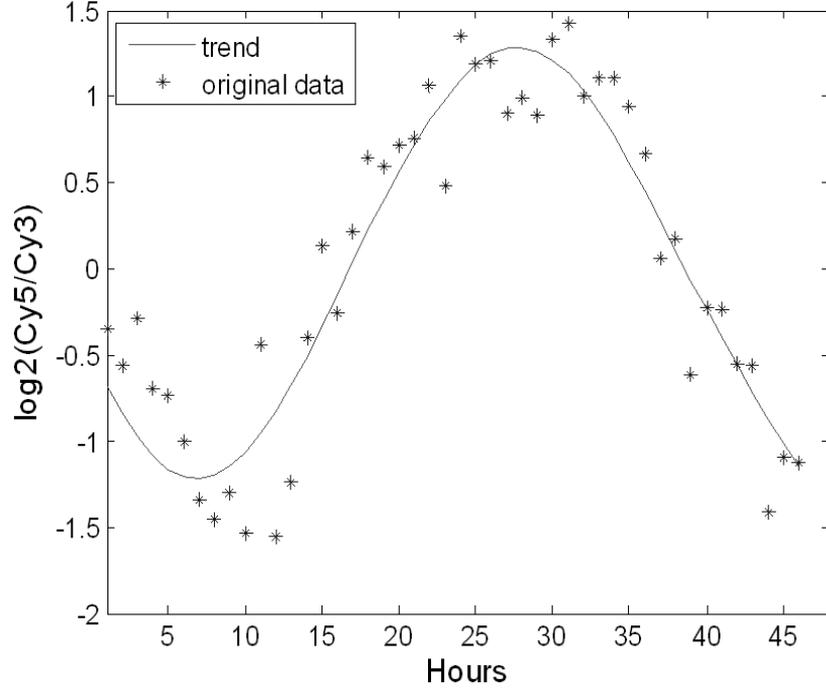
$$E = \frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^M \lambda_i} \quad (1)$$

where  $d$  is the number of leading singular values.

For the majority of transcriptome of IDC, the leading two singular values of expression profiles contain most of the energy and correspond to the signal (Figure 1). Therefore, we group the leading two singular values to reconstruct the expression profiles (Liu *et al.*, 2003) [8]. An example of trend extraction using the SSA is shown in Figure 2.



**Figure 1** The singular values of the data matrix for Dihydrofolate Reductase-Thymidylate Synthase (DHFS-TS).



**Figure 2** Signal extraction of DHFS-TS using the SSA. The leading two singular values which represent 90% energy are chosen for the reconstruction. The window length is  $M = 23$

### 2.3 The AR model based spectral estimation

Periodicity in genome-wide gene expression datasets has been widely used to identify cell-cycle-regulated genes. Being a popular technique for time series analysis, power spectrum estimation has often been used for periodicity identification. If the data are highly periodic, the calculated power spectrum would show sharp peaks at the corresponding frequency points. As one of the nonparametric power spectrum estimation methods, the Fourier analysis is relatively simple, and could be readily calculated using the fast Fourier transform (FFT) algorithm. Let the discrete data sequence be  $s(n)$ ,  $0 < n < N - 1$ , the corresponding estimate of the power density spectrum is

$$\hat{P}_{ss}(\omega) = T \sum_{l=-(N-1)}^{N-1} \hat{r}_{ss}(l) e^{-j\omega l T} \quad (2)$$

where  $T$  is the sampling interval,  $r_{ss}$  and  $N$  represent the autocorrelation sequence and the number of data samples of signal  $s$ , respectively. However, the gene expression data series are often short and noisy, and the spectral resolution from the Fourier-based power spectrum estimation would be seriously degraded due to the well-known windowing effect, which is caused by the unrealistic assumption of nonparametric methods that the autocorrelation estimate  $r_{ss}(l)$  is zero for  $l > N$ .

The AR model for the time series  $s(n)$  is given by

$$s(n) = -\sum_{p=1}^P a_p s(n-p) + u(n) \quad (3)$$

where  $a_p$  are the AR coefficients,  $P$  is the order of the AR model, and  $u(n)$  is a white noise sequence. The forward linear prediction estimation is given by

$$\hat{s}(n) = -\sum_{p=1}^P a_p s(n-p) \quad (4)$$

where  $\hat{s}$  is used to denote an estimate of  $s$ . Therefore, the AR model can be interpreted as the estimation at current time index  $n$  based on  $P$  past samples, while  $u(n)$  is the estimation error.

Note that the AR method is a generative method, where the signal generation model is given by Equation (4). Hence, AR modeling can extrapolate the values of the autocorrelation for  $l > N$ . The parameter  $P$  is called the order of the AR model and it can be estimated from the correlation matrix of the observed data. The corresponding power spectrum density can be estimated as follows,

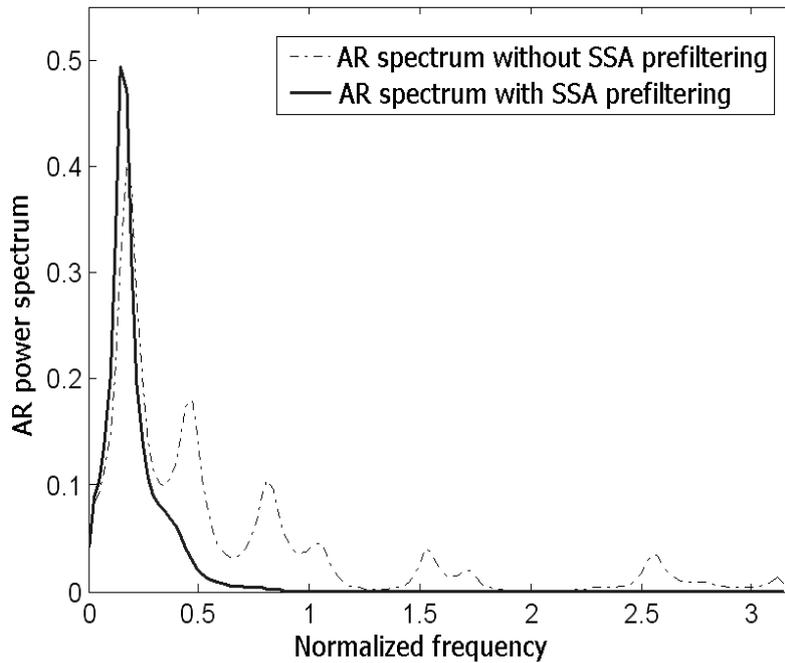
$$\begin{aligned} P_{ss}(\omega) &= T \sum_{l=-\infty}^{\infty} r_{ss}(l) e^{-jl\omega T} \\ &= \frac{T\sigma^2}{\left| 1 + \sum_{p=1}^P a_p e^{-j2\pi\omega p T} \right|^2} \end{aligned} \quad (5)$$

where  $P_{ss}(\omega)$  represents the power spectral density of the signal  $s$  at frequency  $\omega$ . Equation (5) shows that the region of support of the autocorrelation  $r_{ss}$  is  $(-\infty, \infty)$ . For this reason, the AR power spectral density estimators do not possess the sidelobe phenomenon of the classic spectral estimators and can yield higher frequency resolution than nonparametric methods (Marple 1987; Yan 2002; Yan *et al.*, 2007; Yeung *et al.*, 2004) [9,16,17,18]. There are a number of algorithms developed for estimating the parameters  $a_k$ . In this work, we adopt the Yule-Walker method to estimate these parameters (Yan 2002; Yeung *et al.*, 2004) [16,18]. The Yule-Walker method, also called the autocorrelation or windowed method, computes the AR parameters by forming a biased estimate of the signal's autocorrelation function, and solving the least squares minimization of the forward prediction error (Marple 1987) [9].

#### *2.4 The combination of SSA and AR modeling*

The major advantage of the AR model based power spectrum over classical Fourier analysis is its high frequency resolution obtained by fitting a relatively high order AR model to the original data sequence. However, the AR spectrum is also sensitive to noise. When the signal to noise ratio is low, the accuracy of the parameter estimation in Equation (3) would be reduced substantially. A higher order AR model has to be used to improve the frequency resolution, but the usage of higher order would induce the appearance of spurious peaks. According to Keppenne and Penland (Keppenne *et al.*,

1992; Penland *et al.*, 1991) [5,10], SSA can be used as a data-adaptive noise filter. By applying AR power spectrum estimation to the SSA pre-filtered data series, noise-free or noise-reduced frequency spectrum could be obtained. As shown in Figure 3, due to the removal of noise using the SSA, spurious peaks are eliminated from the spectrum. We should point out here that SSA and AR modeling are general signal processing methods and can be applied to the analysis of many types of time series data. There are also software packages available to implement these algorithms. For an example of the software, see <http://www.systat.com/>.



**Figure 3** The AR spectra of the expression profile of DHFR-TS with and without SSA filtering

### 2.5 The periodicity detection using the AR model

Assuming that the AR spectrum reaches its peak point at the frequency  $f_i$  and considering the frequency band  $[f_{i-1}, f_{i+1}]$  as  $f_i$ 's region of influence (ROI). We take the ratio of the power in  $f_i$ 's ROI to the total power of the signal to quantify the periodicity of the expression profile of each gene:

$$S = \frac{power_i}{power_{total}} \quad (6)$$

where  $power_i$  and  $power_{total}$  represent the power over  $f_i$ 's ROI and the total power of the signal, respectively.

First, each expression data is reconstructed using SSA. Only the expression profiles with the singular value ratios in Equation (1) ( $d = 2$ ) greater than 0.6 are to be

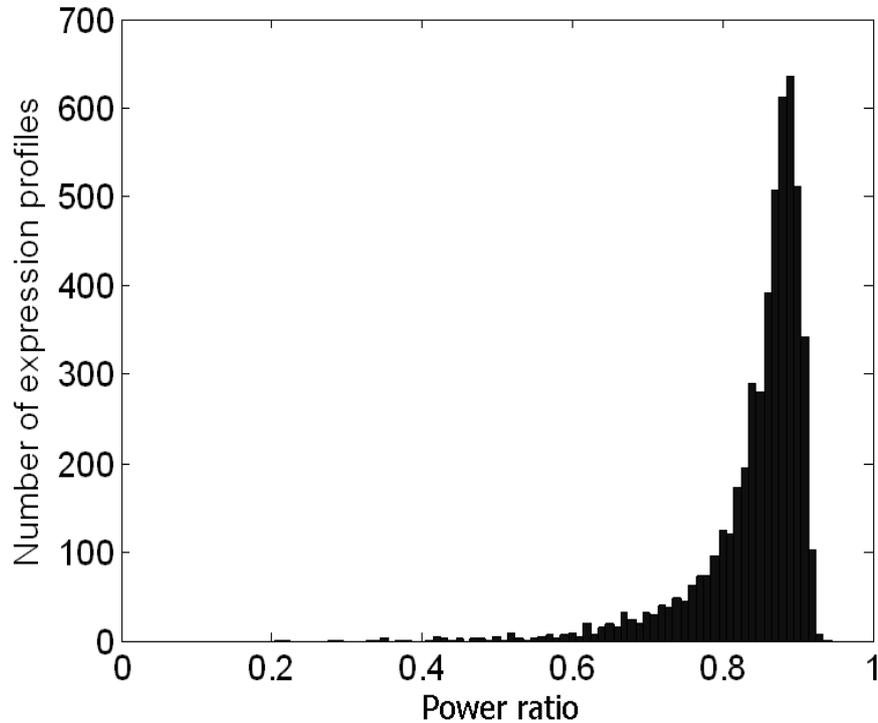
## *Spectral Analysis of Microarray Gene Expression Time Series Data*

reconstructed. Second, the AR spectrum is calculated for each reconstructed expression profiles. Then the frequency  $f_i$  at peak value point and the ratio of the power in  $f_i$ 's ROI to the total power are calculated according to Equation (6). Finally, the profiles are screened according to following rule: if the power ratio calculated previously is larger than 0.7 (which is the same value as that used in Bozdech *et al.* (2003) [1]), the corresponding profile would be selected as periodic, otherwise, we consider it lacking of periodicity and discard it.

### **3 Results**

#### *3.1 Periodic profile detection*

We have tested our algorithm on the expression data of the IDC of *P. falciparum*. There are a total of 5080 expression profiles in the dataset. In DNA microarray data analysis, one of the major challenges is to dissociate actual gene expression values from noise effectively. We used the SSA for trend estimation and have only selected those profiles whose dominant components contain most of the energy for further analysis. Then we performed spectrum analysis of the reconstructed expression profiles by using AR spectral estimation. When an expression profile is highly periodic, its power spectrum would have sharp peaks at the corresponding frequency points. We utilized the power ratio to filter the reconstructed expression profiles. The histogram of the power ratio of the expression profiles of the IDC data is shown in Figure 4. By using our periodicity detection method, 4496 periodic profiles are found to be periodic. Compared with the Bozdech *et al.* (2003) [1]'s result, an additional 777 periodic oligonucleotides are detected using our algorithm.



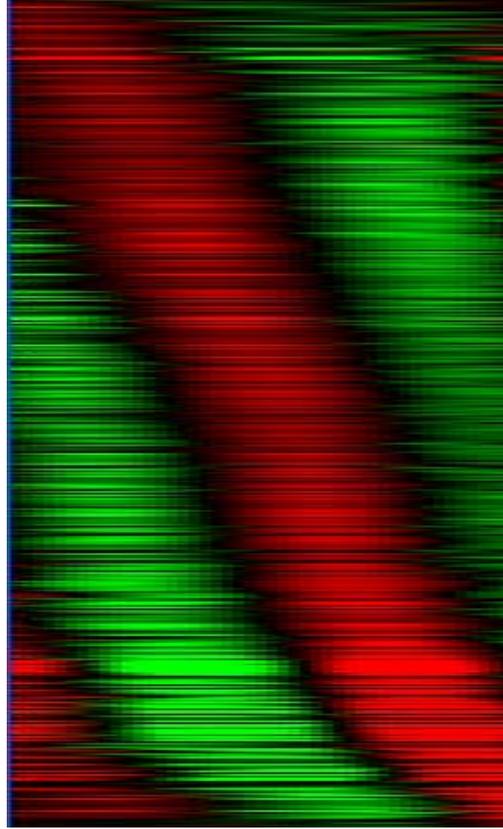
**Figure 4** The histogram of the power ratios of the express profiles of the intraerythrocytic developmental cycle

### 3.2 Analysis of classification result

It is generally accepted that the function of a gene is related to the initial phase of its expression profile (Bozdech *et al.*, 2003; Rustici *et al.*, 2004) [1,12]. So for the convenience of gene function classification, we have ordered the expression profiles of the extracted 4496 oligonucleotides according to their peak time points of expression profiles. Figure 5 shows a continuous cascade of gene expressions, which correspond to the developmental stages throughout the IDC, that is, ring, trophozoite and schizont stages. According to the sharp transitions of ring-to-trophozoite (at the 17h time point), trophozoite-to-schizont (at the 29h time point) and schizont-to-rings stages (at the 45h time point) (Bozdech *et al.*, 2003) [1], 4496 selected genes could be categorized into four stages on the basis of the peak time points of expression profiles in Figure 5. The comparison of the classification results of the oligonucleotides assigned to these stages by our method with those in Bozdech *et al.* (2003) [1] is shown in Table 1. Bozdech *et al.* (2003) [1] has listed all possible functional gene of *P. falciparum*. Among these functional genes, some cannot be detected using the method in Bozdech *et al.* (2003) [1] due to the low power ratio score. In contrast, our method using the SSA helps to smooth out noise and makes the dominant spectral component much more evident. In addition, the AR spectrum produces higher frequency resolution than classical Fourier spectrum. By combining advantages of SSA and AR models, periodicity in the gene expression profiles is extracted much more accurately. More functional genes are detected compared

### *Spectral Analysis of Microarray Gene Expression Time Series Data*

with the Fourier analysis technique used in Bozdech *et al.* (2003) [1]. In Table 2, we provide the list of functional genes identified by our algorithm that was not found by the method in Bozdech *et al.* (2003) [1].



**Figure 5** The phaseogram of the transcriptome of the IDC of *P. falciparum*. 4496 genes are ordered along the y axis in the order of the time of their peak expression

**Table 1** Comparison of the classification results of oligonucleotides in three different stages. More genes can be identified using our method

Stages	Our method	Method in Bozdech <i>et al.</i> (2003) [1]
Ring/early Trophozoite	1970	1563
Trophozoite/early Schizont	1524	1296
Schizont	709	625
Early ring	293	235

**Table 2** The list of genes for different functional groups of *P. falciparum* detected by our method that was not found by the method in Bozdech *et al.* (2003) [1]

Oligo-nucleotide	Gene ID	Description	Power ratio	Functional groups
b471	PFB0715w	DNA-directed RNA polymerase II second largest subunit	0.88182	Transcription machinery
opff72413	MAL6P1.189	Hexokinase	0.84831	Glycolytic pathway
j22_5	PF10_0086	Adenylate kinase	0.84539	Ribonucleotide synthesis
c345	PFC0520w	26S proteasome regulatory subunit S14	0.7706	Proteasome
F20448_1	MAL8P1.142	proteasome beta-subunit	0.73066	Proteasome
I14413_3	PFI0630w	26S proteasome regulatory subunit	0.72985	Proteasome
j110_4	PF10_0174	26s proteasome subunit p55	0.76564	Proteasome
j64_2	PF10_0298	26S proteasome subunit	0.79793	Proteasome
j73_12	PF10_0081	26S proteasome regulatory subunit 4	0.76011	Proteasome
kn3744_1	PF10_0174	26s proteasome subunit p55	0.76345	Proteasome
ks101_10	NULL	N/A	0.83138	Proteasome
M21665_2	PF13_0063	26S proteasome regulatory subunit 7	0.77563	Proteasome
n135_24	PF14_0632	26S proteasome subunit	0.72119	Proteasome
n155_11	PF14_0025	proteosome subunit	0.73009	Proteasome

Genes expressed during the mid to late schizont and early-ring stage encode proteins predominantly involved in highly parasite-specific functions facilitating various steps of host cell invasion. The highly parasite-specific functions implied that they should serve as good targets for both drug discovery and vaccine-based antimalarial strategies (Bozdech *et al.* 2003) [1]. The additionally identified genes are distributed in the 777 additional oligonucleotides found by our method and they would be useful for the future identification of novel targets for anti-malarial therapies. In general, there are two gene targets to be considered for new drug discovery: apicoplast-targeted genes and proteases. In Table 3, we provide a comparison of the number of the two gene targets detected by our method and by Bozdech *et al.* (2003) [1].

**Table 3** The comparison of the numbers of the two types of potential gene targets for drug discovery detected by our method and the method in Bozdech *et al.* (2003) [1]

Potential genes targets	Our method	The method in Bozdech <i>et al.</i> (2003)
Apicoplast-targeted	409	358
proteases	115	88

## *Spectral Analysis of Microarray Gene Expression Time Series Data*

Merozoite invasion is one of the most promising target areas for antimalarial vaccine development. Among these invasion proteins are seven of the well known malaria vaccine candidates, including Apical Merozoite Antigen-1 (AMA1), Merozoite Surface Protein-1 (MSP1), Merozoite Surface Protein-3 (MSP3), Merozoite Surface Protein-5 (MSP5), Erythrocyte Binding Antigen-175 (EBA175), Rhoptry-Associated Protein-1 (RAP1) and Ring-infected Erythrocyte Surface Antigen-1 (RESA1). It is now widely accepted that genes with the same or similar function are likely to have similar expression profiles. Therefore, we have used the similarity to identify genes with possible involvement in the merozoite invasion process. The similarity of 4496 expression profiles to seven known vaccine candidates is evaluated using the Euclidian distance. There are a total of 267 genes, constituting the top 6% of the detected periodic genes in the IDC, with minimum distance to these seven genes (the same distance threshold is adopted from Bozdech *et al.* (2003) [1]). Five additional genes are detected by our method, which are listed in Table 4. They are all from the late schizont stage.

**Table 4** The 5 additionally detected genes as potential antimalarial vaccine candidate

Oligonucleotides	Genes	Description
n139_5	PF14_0757	Hypothetical protein
n143_54	PF14_0183	RNA helicase
n168_2	PF14_0530	Ferlin, putative
opfblob0035	PFD0430c	Hypothetical protein
opfl0111	PFL2110c	Hypothetical protein

## **4 Conclusion**

Spectral analysis of the transcriptome of the asexual intraerythrocytic development cycle (IDC) of *P. falciparum* offers an effective means for the studies of cyclic gene expression and future drug and vaccine targets. In this paper, we have proposed a new scheme based on SSA and AR power spectral analysis for the detection of periodically expressed genes in the IDC of *P. falciparum*. SSA allows the dominant trend to be extracted from the noisy expression profiles, and subsequent AR spectrum analysis provides higher frequency resolution in the periodicity detection. We have detected more periodically expressed profiles in the *P. falciparum* dataset. Compare with the results of Bozdech *et al.* (2003), we have obtained 14 additional functional genes undetected by them. That is, we have detected more gene targets for new drug discovery, and we have identified 5 additional genes that are potential antimalarial vaccine candidate. The result shows that our method can not only detect more periodic genes but also can find genes that could be useful targets for potential drug/vaccine design.

## Acknowledgments

This work is supported by a grant from the Hong Kong Research Grant Council (project CityU 122607). Liping Du is also supported by the National Natural Science Foundation of China (No. 60773074) and the National High Technology Research and Development Program of China (No. 2007AA01Z213).

## References

1. Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J.C., and DeRisi, J.L. (2003) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *Plos Biology*, 1, 1-16
2. Du, L., Wu, S., Liew, A.W.C., Smith, D., and Yan, H. (2006) Parametric spectral analysis of malaria gene expression time series data. Proc. 2<sup>nd</sup> International Symposium on Computational Life Sciences, Lecture Notes in Bioinformatics. M. Berthold, R. Glen and I. Fischer (eds.), Springer-Verlag. **4216**, 32-41
3. Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., and Bowman, S., et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511.
4. Golyandina, N., Nekrutkin, V., and Zhigljavsky, A. (2001) Analysis of time series structure: SSA and related techniques. Chapman & Hall/CRC, Boca Raton, Florida.
5. Keppenne, C.L., and Ghil, M. (1992) Adaptive filtering and prediction of the Southern Oscillation index. *J. Geophys. Res.* **97**, 20449-20454.
6. Le Roch, K.G., Zhou, Y.Y., Blair, P.L., Grainger, M., Moch, J. K., Haynes, J. D., De la Vega, P., Holder, A.A., Batalov, S., and Carucci, D. J. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301**, 1503-1508.
7. Liew, A.W.C., Xian, J., Wu, S., Smith, D., and Yan, H. (2007) Spectral estimation in unevenly sampled space of periodically expressed microarray time series data. *BMC Bioinformatics* **8**, 137:1-19
8. Liu, L., Hawkins, D.M., Ghosh, S., and Yong, S.S. (2003) Robust singular value decomposition analysis of microarray data. *PNAS* **100**, 13167-13172.
9. Marple, S. L., (1987) Digital spectral analysis: with applications, Prentice-Hall, Englewood Cliffs, New Jersey.
10. Penland, C., Ghil, M., and Weickmann, K.M. (1991) Adaptive filtering and maximum entropy spectra with application to changes in atmospheric angular momentum. *J. Geophys. Res.* **96**, 22659-22671.
11. Rajarajeswari, B., Eyke, H., Nils, W., and Jörg, K. (2005) Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics* **21**, 1069-1077.
12. Rustici, G., Mata, J., Kivinen, K., Lió, P., Penkett, C.J., Burns, G., Hayles, J., Brazma, A., Nurse, P., and Bähler, J. (2004) Periodic gene expression program of the fission yeast cell cycle. *Nature Genetics* **36**, 809-817.
13. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, 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.
14. Vautard, R., and Ghil, M. (1989) Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Physica D.* **35**, 395-424.
15. Vautard, R., Yiou, P., and Ghil, M. (1992) Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D.* **58**, 95-126.

*Spectral Analysis of Microarray Gene Expression Time Series Data*

16. Yan, H. (ed.), (2002) *Signal Processing for Magnetic Resonance Imaging and Spectroscopy*, Marcel Dekker, New York.
17. Yan, H., and Pham, T. (2007) Spectral estimation techniques for DNA sequence and microarray data analysis. *Current Bioinformatics* **2**, 145-156.
18. Yeung, L.K., Szeto, L.K., Liew, A.W.C., and Yan, H. (2004) Dominant spectral component analysis for transcriptional regulations using microarray time series data. *Bioinformatics* **20**, 742-749.