Wavelet analysis on temporal expression profiles of malaria parasite

Hong Cai[1] Sos Agaian[1] Maribel Sanchez[3] Jianying Gu[2] Yufeng Wang[3]

[1] Department of Electrical Engineering, University of Texas at San Antonio
San Antonio, TX 78249, USA
[2] Department of Ecology and Evolution, University of Chicago
Chicago, IL 60637, USA
[3] Department of Biology, University of Texas at San Antonio
San Antonio, TX 78249, USA.
Correspondence: ywang@utsa.edu

Abstract

It is important to monitor the temporal behavior of gene expression for achieving a systems level understanding of cellular and developmental networks. The difficulty in temporal microarray analysis stems from the high dimensionality of data. In this paper, we present a time-frequency wavelet analysis on time-course microarray data of malaria parasite, which include the expression profiles of over 4,400 genes during a 48-hour developmental cycle in the red blood cell. This approach is able to extract features in both time and frequency domains and reduce the data dimensionality. The classification based on low-dimensional feature vectors allows for an enhanced detection of time-dependent expression patterns and lead to the prediction of putative genes with specific functionality.

Keywords: Microarray, time course, wavelet, malaria, gene networks.

1. Introduction

With the advent of microarray technology, cell division, differentiation, programmed cell death (apoptosis), and other cellular processes can be monitored by expression profiles of thousands of genes along a time course [1-3]. The main premise of this high-throughput transcriptional analysis is that complex cellular phenomena can best be understood by observing many key time-specific cascade events. One such example involves a fine coordination between evolutionarily conserved, but highly specific, families of proteins to regulate yeast cell cycles [1, 2].

It is computationally challenging to uncover dynamic gene networks from time-series microarray data, due to complex sources of noise, large systematic or random variations and high dimensionality [4, 5]. For example, in attempts to use time course data to build transcriptional models, the dimensionality problem arises from the discrepancy between the large number of genes being measured and the relatively small number of time points being used. Current analysis approaches are mainly focused on classification of co-expressed genes, using supervised methods such as support vector machine [6], artificial neural networks [7], or unsupervised clustering methods such as hierarchical clustering [8], K-means clustering [9], and self-organizing map [10]. Methods that specifically deal with oscillating time-series data include Fourier analysis [2, 3], wavelet decomposition [11], definition of selection threshold [12], and shape-invariant statistical model [13].

In this paper, we apply an approach that combines time-frequency wavelet decomposition analysis and higher-order statistics to extract the features of microarray data of malaria parasite in a 48-hour red blood cell cycle. We classify these features by Kernel Fisher Discriminant (KFD), which yield a set of genes that show characteristic expression patterns over the developmental cycle.

2. Data and Methods

2.1. Time course data set

The data set used in this study includes microarray expression profiles of malaria parasite *Plasmodium falciparum*, at one-hour time intervals during the course of the 48-hour developmental cycle [3] (http://malaria.ucsf.edu/SupplementalData.php). Each chip consists of a total of 7,462 probes that represent over 4,488 genes. The final data set after quality control filtering and normalization contained the

profiles of 46 time points excluding 23-hour and 29-hour time points. The signals used for further time-frequency analysis were the log2(Cy5/Cy3) ratios, in which Cy5 signals corresponded to individual synchronized time points, and Cy3 signals corresponded to reference asynchronized samples.

2.2. Feature extraction by wavelet analysis

The goal of our analysis is to develop methods to extract features that best represent the behavior of gene networks in a time course.

In the original paper, Bozdech et al. [3] extracted phase information using Fourier Transform (FT). FT decomposes a function into a sum of sinusoidal waves. These sinusoids are very well localized in the frequency domain, but not in the time domain. This drawback is minor only if the signal properties do not change much over time. However, most interesting biological signals contain numerous nonstationary or transitory characteristics that indicate drift, trends, abrupt changes, breakdown points, and beginnings and ends of events. These characteristics are obscured from Fourier analysis.

To capture the properties of gene expression in both time and frequency domains, we employed wavelet transform. The wavelet transform can use the long time intervals with more precise low frequency information, and shorter regions with high-frequency information.

The wavelet transform with multilevel structures can be viewed as decomposition by high-pass and low-pass filter banks. After wavelet decomposition, two higher-order statistics, skewness and kurtosis, can be employed to analyze wavelet coefficients and to build feature vectors [14] (Fig. 1).

We derive feature vectors for 12 functional classes of proteins presented in Bozdech et al. that showed distinct developmental profiles [3].

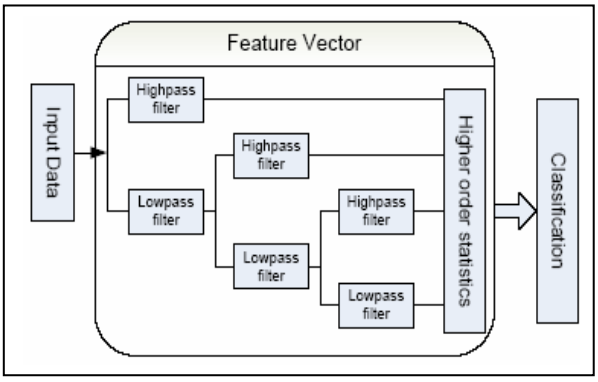

Fig. 1 Wavelet based feature vector

2.3. Classification based on feature vectors

The nonlinear classification, Kernel Fisher Discriminant (KFD) is employed to separate feature vectors of different functional classes.

Our training set included 12 functional classes of 530 microarray probes [5]. The test data set consisted of over 6800 probes excluding the training data. Function scalar values obtained from training can discriminate a known or unknown gene characteristic.

3. Results and Discussion

3.1. Gene properties were captured by wavelet based feature vectors

We used Daubechies wavelet (db8) to conduct three-level decompositions. Approximation coefficients and detail coefficients were obtained by low-pass filter and high-pass filter, respectively. Fig. 2 and Fig. 3 show the approximation coefficients and detail coefficients of multilevel decompositions for two functional classes, transcription which is essential for parasite development and meroziote invasion which is crucial for parasite infection. Clearly, two types of wavelet coefficients hold different information of original signals; moreover, the two functional classes display distinct wavelet properties.

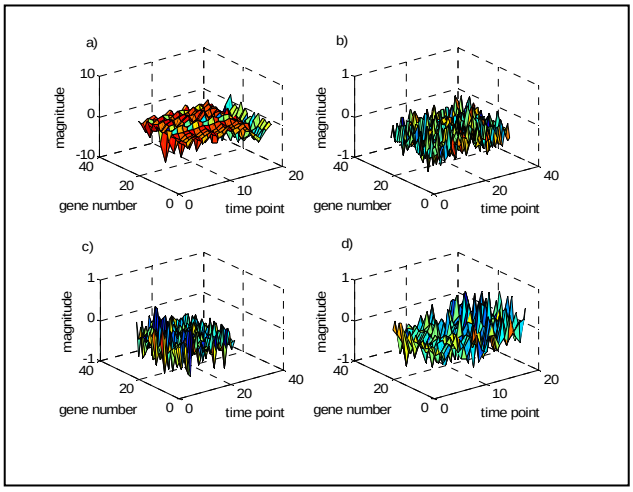

Fig. 2. Multi-level wavelet decomposition for 23 genes that belong to transcriptional machinery: (a) Approximation coefficients. (b) First level detail coefficients. (c) Second level detail coefficients. (d) Third level detail coefficients.

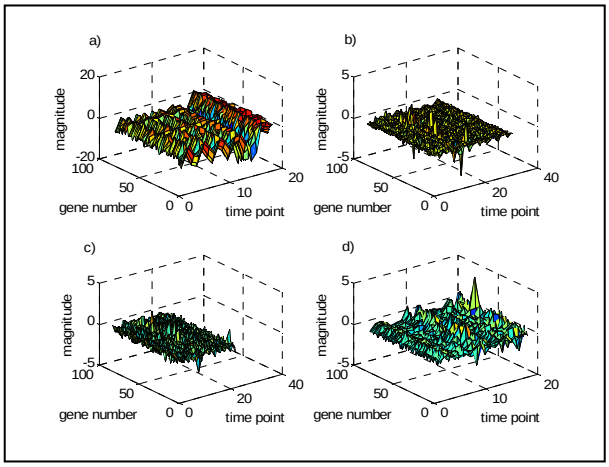

Fig. 3. Multi-level wavelet decomposition for 87 genes that are involved in merozoite invasion: (a) Approximation coefficients. (b) First level detail coefficients. (c) Second level detail coefficients. (d) Third level detail coefficients.

Mean, variance of approximation coefficients and skewness and kurtosis of multi-level detail coefficients were used to build feature vectors for specific functional classes.

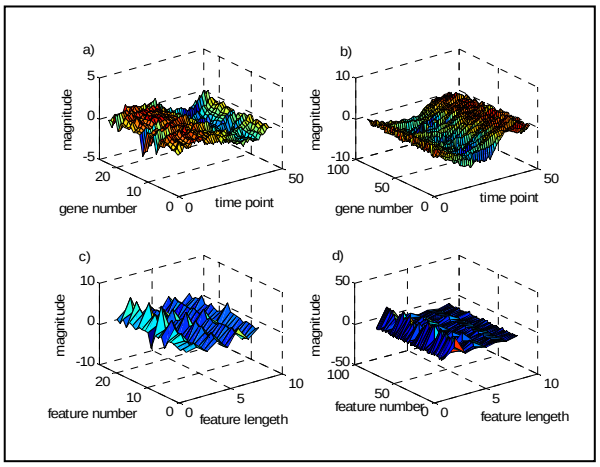

Fig. 4. Gene properties captured by feature vectors: (a) Original data of "transcriptional" functional class. (b) Original data of "merozoite invasion" functional class. (c) Feature vectors of "transcriptional" functional class. (d) Feature vectors of "merozoite invasion" functional class.

Compared to the original data (Fig. 4a and 4b), the dimension of corresponding feature vectors (Fig. 4c and 4d) was reduced from 46 (time points) to 8, which made further computation cost-effective. Furthermore, KFD based on these feature vectors separated 23 and 87 genes in these two functional classes with 99% accuracy, suggesting that the time-dependent gene expression properties are well represented by the low-dimensional feature vectors.

3.2. KFD classification identified new genes that share similar temporal expression profile

Using 23 putative genes in transcriptional machinery as a training set, the KFD analysis on over 6800 probes yielded a total of 236 positive hits. Sharing similar developmental profiles which show peaks in ring and early-trophozoite stages, these positive probes can be grouped into three categories:

(1) Genes involved in transcriptional process: as shown in Table 1, multiple probes that correspond to DNA-directed RNA polymerase II (PFC0805w) were picked. In addition, several putative transcription factors with characteristic domains such as zinc finger domain may play a role in transcriptional regulation.

Table 1: Putative genes detected by wavelet analysis and KFD that may be involved in transcription.

Oligo_ID	Gene_ID	Annotation
f22770_1	PFC0805w	DNA-directed RNA pol II
opfi17677	PFC0805w	DNA-directed RNA pol II
opfc0750	PFC0805w	DNA-directed RNA pol II
j132_12	PF10_0327	Myb2,Transcription factor
opfn0273	PF14_0241	basic transcription factor 3b
f21506_2	MAL8P1.131	Transcription factor Gas41 homologue
n134_51	PF14_0612	Hypothetical protein with putative zinc finger domain
f34582_1	MAL6P1.193	transcription factor-like, Zn-finger C2HC domain

(2) Genes involved in processes that are tightly associated with transcription. Table 2 lists several representative genes that may be components of downstream processes: (a) pre-RNA processing after transcription, which requires a complex of small nuclear proteins (snRNPs) and splicing factors. (b) Translation initiation which requires eukaryotic initiation factors and helicases.

(3) Genes that encode hypothetical proteins.

One of the major limitations on the use of genomic data to better understand infectious diseases is our inability to assign a functional identity to a large fraction of the recognized open reading frames in a given genome. Nowhere is this more problematic than in the case of malaria parasite; in this organism 60% of the open reading frames are annotated as "hypothetical". Classifying these hypothetical proteins is very important for malaria research: because the life cycle of malaria parasite is dynamic and complex,

encompassing infections of mosquito and human hosts, the first step toward understanding a hypothetical protein is to characterize when and where it is expressed. For example, a portion of hypothetical proteins that are classified into “transcription” may represent unknown transcription factors, which modulate parasite-specific gene regulatory networks (Wang et al. unpublished data).

Table 2: Putative genes detected by wavelet analysis and KFD that may be involved in biological processes related to transcription.

Oligo_ID	Gene_ID	Annotation
D49176_31	PFD0265w	Splicing factor
M18079_5	MAL13P1.120	Splicing factor
E29750_1	PFE0160c	Splicing factor
F38861_1	MAL6P1.104	Step II splicing factor
I5180_2	PFI0475w	snRNP
M28007_3	MAL13P1.35	U1 snRNP
oPFH0006	MAL8P1.48	snRNP
d16785_20	PFD1070w	eukaryotic initiation factor
L2_62	PFL0335c	eukaryotic initiation factor 5
M45177_10	PF13_0178	eukaryotic initiation factor 6
N150_48	PF14_0104	eukaryotic initiation factor 2
D49176_14	PFD0245c	RNA helicase
B312	PFB0445c	Putative helicase
E19278_2	PFE1085w	DEAD-box subfamily ATP-dependant helicase

In conclusion, our method that combines wavelet-based higher order statistics and KFD provides an effective means to discover novel genes from time-series microarray profiles and brings new insight into the regulatory gene networks.

This study is supported by San Antonio Area Foundation, NIH RCMI grant 2 G12 RR013646-06A1, and a UTSA start-up fund to Y. Wang.

4. References

[1] 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, and R. W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, pp. 65-73, 1998.

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

[3] Z. Bozdech, M. Llinas, B. L. Pulliam, E. D. Wong, J. C. Zhu, and J. L. DeRisi, "The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum," *Plos Biology*, vol. 1, pp. 85-100, 2003.

[4] J. Quackenbush, "Computational analysis of microarray data," *Nature Reviews Genetics*, vol. 2, pp. 418-427, 2001.

[5] Y. F. Leung and D. Cavalieri, "Fundamentals of cDNA microarray data analysis," *Trends in Genetics*, vol. 19, pp. 649-659, 2003.

[6] Y. Lee and C. K. Lee, "Classification of multiple cancer types by tip multicategory support vector machines using gene expression data," *Bioinformatics*, vol. 19, pp. 1132-1139, 2003.

[7] J. Vohradsky, "Neural model of the genetic network," *Journal of Biological Chemistry*, vol. 276, pp. 36168-36173, 2001.

[8] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, pp. 14863-14868, 1998.

[9] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nature Genetics*, vol. 22, pp. 281-285, 1999.

[10] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, pp. 2907-2912, 1999.

[11] R. R. Klevecz and H. B. Dowse, "Tuning in the transcriptome: basins of attraction in the yeast cell cycle," *Cell Proliferation*, vol. 33, pp. 209-218, 2000.

[12] K. F. Storch, O. Lipan, I. Leykin, N. Viswanathan, F. C. Davis, W. H. Wong, and C. J. Weitz, "Extensive and divergent circadian gene expression in liver and heart," *Nature*, vol. 417, pp. 78-83, 2002.

[13] Y. Luan and H. Li, "Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data," *Bioinformatics*, vol. 20, pp. 332-339, 2004.

[14] H. Farid, "Detecting hidden messages usinh higher-order statistical models.," presented at International Conference on Image Processing (ICIP), Rochester, NY, 2002.