Proceedings of the 2006 WSEAS International Conference on Mathematical Biology and Ecology, Miami, Florida, USA, January 18-20, 2006 (pp128-133)

Exploring Malaria Developmental Expression Profiles Using Wavelet Analysis and Support Vector Machine

HONG CAI[1], SOS S. AGAIAN[1], MARIBEL SANCHEZ[2], YUFENG WANG[2*]
[1] Department of Electrical Engineering, [2] Department of Biology
University of Texas at San Antonio
San Antonio, TX 78249
USA
[*]Correspondence: yufeng.wang@utsa.edu

Abstract: - As a new central platform in functional genomics, high-throughput microarray technology allows a systems-level investigation of fundamental biological processes. Among various microarray analyses, time-series analysis presents a computational challenge due to the intrinsic nature of high dimensionality. In this paper we propose a novel approach to combine wavelet analysis with support vector machine learning scheme. This approach is able to extract gene features in both time and frequency domains and to reduce dimensionality. The data set includes the expression profiles of over 4,000 genes in malaria parasite *Plasmodium falciparum*, during a 48-hour intraerythrocytic developmental cycle. After wavelet decomposition, higher-order statistics are employed to analyze wavelet coefficients and to build feature vectors. Our preliminary analysis of 14 functional classes shows that transcriptional regulation and developmental progression are highly correlated. Novel malarial genes have been identified based on the tempo-specific patterns and further bioinformatics data-mining. Classifying novel "hypothetical proteins" to network modules enables a targeted functional characterization, as for a parasite with multiple hosts and a dynamic life cycle, "when and where" to initiate wet-lab experiments is of critical importance.

Key-Words: - malaria, microarray, time series, wavelet, SVM, systems biology, gene networks

1 Introduction

Malaria is one of the most devastating infectious diseases. Over 500 million cases are reported and about 2 million people die annually. Four protozoan parasite species of the genus *Plasmodium*, *P. falciparum*, *P. vivax*, *P. malariae*, and *P. ovale*, are known to cause the disease in humans.

Although useful therapies against malaria have existed for over 400 years, resistant parasite populations have appeared. The factors including the rapid spread of drug resistance, the lack of an effective vaccine, and the steadily rising resistance of the mosquito vectors to insecticides have led to an urgent need for new antimalarial strategies.

The complete genome of *P. falciparum* was released in 2002 and genomes of other *Plasmodium* species and of other *Apicomplexans* have been released more recently [1-3]. This has opened the door to identify new drug and vaccine targets. However, difficulties stemming from our inability to assign functionality to over 60% of the putative genes slow the genome-based target discovery [1].

A solution to this problem lies in the domain of systems biology, which envisions a high-level view of an organism, thereby offering a better understanding of cellular networks and interactions among network components. A network view would allow us to build models of how the *Plasmodium* parasite functions – the protocols that guide the system, the modules that comprise the system and so on [4-6]. Significantly, *a priori* information as to the identity or function of a gene is not necessary for the gene to be placed in a network, but, such a gene can be targeted for further study as a therapeutic target if it proves to play a key role in the network.

Microarray technology has enabled a systems biology approach to study temporal specific gene networks, by monitoring the transcriptional profiles of genes in a time-series manner [7, 8]. A ground-breaking work by Bozdech et al. [7] examined the expression profiles of *P. falciparum* every hour for the entire duration of the blood stage (48 hours) and presented a blueprint of transcriptomes.

It is an unprecedented challenge to uncover dynamic genetic regulatory networks (GRNs) from time-series microarray data, due to the complex sources of noise and variations, and high dimensionality. Various methods have been employed to classify co-expressed genes, using supervised learning such as support vector machine (SVM) [9, 10], neural networks [11], or unsupervised clustering methods such as

hierarchical clustering [12] and K-means clustering [13]. Alternatively, probabilistic graphical modeling methods such as Boolean networks [14] and Bayesian networks [15] have been developed to infer GRNs.

Special care must be taken to account for the oscillating nature in time-series data. Recently, we have developed a linear-Gaussian state-space model and variational Bayes Expectation Maximization algorithm which takes temporal correlation into account to infer the yeast cell cycle network [16].

However, in the case of malaria, direct probabilistic graphical modeling may not be applicable due to our limited knowledge about the network components. In this paper, we propose a pipeline to discover novel network component that combines wavelet analysis for time-expression correlation, with powerful supervised learning using SVM, followed by genomic data mining.

This paper is organized as follows: Section 2 describes the microarray dataset and the analysis pipeline. In Section 3, we present the experimental results. The conclusions are present in Section 4.

2 Dataset and Methods

2.1 Dataset: time series microarray data

The dataset included the expression profiles of malaria parasite during 48-hour red blood cell cycle (http://malaria.ucsf.edu/SupplementalData.php) [7]. 46 consecutive time points were included except 23-hour and 29-hour during which synchronized samples were not available. Preprocessing and normalization led to a complete dataset of signals for 7092 probes corresponding to over 4000 genes. The $\log_2$(Cy5/Cy3) values were used for study, where Cy5 and Cy3 signals corresponded to the synchronized and asynchonized samples at each time point.

2.2 Wavelet analysis

To reveal the time-expression correlation in developmental processes, we employed wavelet analysis. Fast Fourier Transform (FFT) was used in the original study to extract the phase information [7]. However, as FFT only processes signals in the frequency domain, it is unable to capture important signals with non-stationary characteristics that indicate regulatory trends, sudden changes, breakdown points, and initiations and terminations of cellular events.

Such transitory signals may be captured by wavelet analysis which processes signals in both time and frequency domains. Despite its wide applications in the fields of signal processing [17], wavelet analysis has rarely been employed in microarray data.

A family of wavelets from φ by dilating and translating is given in equation (1). The parameter a controls the scale (or size of details) and then the scale becomes increasingly finer as a approaches 0. Wavelet can be considered as a mathematical microscope due to this property.

$$\varphi^{(a,b)}(x) = |a|^{-1/2}\,\varphi\left(\frac{x-b}{a}\right) \quad (1)$$

φ should satisfy the following admissibility condition as described in equation (2).

$$2\pi\int_R \frac{|\hat{\varphi}(\omega)|^2}{|\omega|}d\omega < \infty \quad (2)$$

Wavelet transform is defined in equation (3). It can be invertible, where $g(t)$ is an input signal.

$$(W_\varphi g)(a,b) = < g, \varphi^{a,b} > = |a|^{-1/2}\int_R g(t)\overline{\varphi\left(\frac{t-b}{a}\right)}dt \quad (3)$$

The wavelet transform with multilevel structures can be viewed as decomposition by high-pass and low-pass filter banks. As shown in Figure 1, a 3-level wavelet decomposition can be achieved by employing a filter bank, where L and H are the analysis low-pass and high-pass filters. Let A_3 be the input to the analysis filter bank. The outputs of the analysis filter bank are then given by

$$A_i(k) = \sum_n L(n-2k)A_{i+1}(n)$$

$$D_i(k) = \sum_n H(n-2k)D_{i+1}(n)$$

where A_i and D_i are defined as the approximation and detail coefficients of the wavelet decomposition of A_{i+1}. After wavelet decomposition, statistics such as mean, variance, higher-order statistics, can be used to analyze wavelet coefficients and to build feature vectors that characterize temporal features. The features contain the approximation coefficients A_0 and the statistical components which are derived from both the approximation and the detail coefficients at each level. These components are defined as follows.

$$\varphi_A = \frac{\mu_{A_2}\mu_{A_1}\mu_{A_0}}{\mu_{A_2}^2+\mu_{A_1}^2+\mu_{A_0}^2},\ \varphi_D = \frac{\mu_{D_2}\mu_{D_1}\mu_{D_0}}{\mu_{D_2}^2+\mu_{D_1}^2+\mu_{D_0}^2}$$

$$\psi_A = \frac{\sigma_{A_2}\sigma_{A_1}\sigma_{A_0}}{\sigma_{A_2}^2+\sigma_{A_1}^2+\sigma_{A_0}^2},\ \psi_D = \frac{\sigma_{D_2}\sigma_{D_1}\sigma_{D_0}}{\sigma_{D_2}^2+\sigma_{D_1}^2+\sigma_{D_0}^2}$$

$$S_A = S(A_i),\ S_D = S(D_i),\ K_A = K(A_i),\ K_D = K(D_i)$$

where μ, σ^2, S and K denote mean, variance, skewness and kurtosis, respectively, and i = 0,1,2.

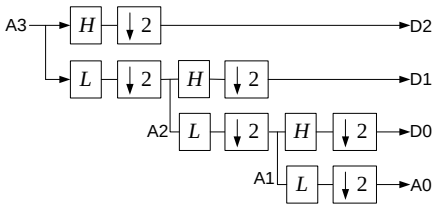

Figure 1. Three-level wavelet transform. The symbol ↓2 denotes the down sampling by 2.

2.3. Classification by SVM

SVM is a powerful supervised learning scheme based on the combination of a feature selection procedure and a classifier. SVM uses a kernel function to define an optimal hyperplane to separate noisy signals into distinct classes.

In the microarray study, SVM was employed to discriminate the feature vectors of various functional classes. Our previous study has shown that SVM with polynomial kernel functions had good performance in classifying yeast benchmark data [18]. However, benchmark data does not exist in malaria. To test SVM, we constructed a learnable dataset based on 14 classes of 530 genes suggested by literature or FFT analysis (Table 1).

When classifying one class, all the genes in this class was labeled positive and the remaining negative. For each class, 2/3 positive genes and 2/3 negative genes were randomly chosen as the training set and the remaining as the testing dataset. This procedure was repeated for 30 times. Each gene can be classified into one of the four: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Because the microarray data are highly imbalanced, i.e., positive instances are much smaller compared to the negative instances, FN is an important feature. Hence we employed three performance measurements:

$precision$ = TP/(TP+ FP), $recall$=TP/(TP+FN),

$f\_measure$=2×($recal$×$precision$)/($recall$+$precision$).

Note that the genes listed in Table 1 only represent a partial set to each class. We further employed SVM classifier to the 6562 (7092-530) microarray probes to identify novel network components. The underlying assumption is that cellular processes are comprised by time-specific cascade events involving co-expressed genes.

2.4 Bioinformatics data mining

Next, we performed genomic analysis on the predicted network components. Their potential functionality was consolidated with the Gene Ontology [19].Conserved domains/motifs in protein sequences were identified by searching the profiles constructed by Hidden Markov Models which are available at the InterPro database (http://www.ebi.ac.uk/interpro/). Multiple alignments were obtained by the program T-coffee (http://www.ch.embnet.org/software/TCoffee.html), followed by manual editing. Graphic presentation of the alignment and consensus sequences were deduced by the program BOXSHADE (http://www.ch.embnet.org/software/BOX_form.html). Phylogenetic trees were inferred by the neighbor-joining method using MEGA 3.1 (http://www.megasoftware.net/). Unweighted maximum parsimony and maximum likelihood were used to consolidate the tree topology. The bootstrap resampling with 1000 pseudoreplicates was carried out to assess support for each individual branch.

3 Experimental Results

3.1 Features extraction by wavelet analysis

Daubechies wavelet (db8) was used at three levels. Figure 2 shows the clear distinction of feature vectors for five functional classes. The three processes, transcription translation, and replication, comprising the central dogma for genetic information flow, take place at consecutive cellular stages. In particular, transcriptional and translation are tightly linked. Their time-expression features, which were not distinguishable by either visual inspection or FFT analysis, can be captured by feature vectors derived from wavelet decomposition (Figure 2). Similarly, wavelet analysis successfully captured temporal distinctions between two consecutive metabolic processes: glycolysis and TCA cycle (Figure 2). Moreover, the dimension of feature vectors was reduced from 46 to 22.

3.2 Classification by SVM

SVM with polynomial kernel functions was employed to classify the malaria "benchmark" data. Table 1 shows the *precision*, *recall*, *f_measure* and their standard deviation of 30 replications for 14 functional classes.

SVM successfully classified genes in classes 2, 8, 9, 10, and 12, as indicated by high *precision*, *recall*, and *f_measure* values. Previously, we showed that PDA, LDA, and SVM usually had low (<20%) *precision*, *recall* and *f_measure* for highly imbalanced data in yeast experiments [7]. Similarly, one class with small size showed 0 *precision*, *recall*, indicating all the positive instances recognized were wrong. Nevertheless, the transcription class, given its small number of positive instances and highly dynamic biology, SVM achieved reasonably good performance.

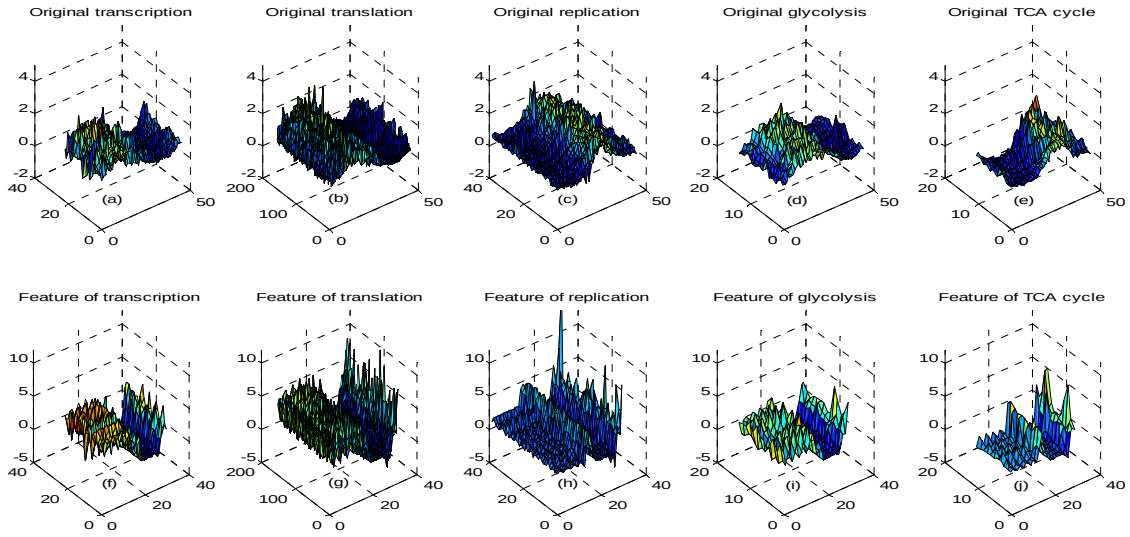

Figure 2. Features of five functional classes extracted by wavelet analysis.

Table 1. SVM classification of 14 functional classes based on wavelet feature vectors

Class	Function (#genes)	*precision* (%)	*recall (%)*	*f_measure* (%)
1	transcription (23)	22.13±11.28	30.86±14.40	25.14±11.64
2	translation (159)	**79.77±4.58**	**81.21±5.73**	**80.29±3.36**
3	Glycolysis (14)	43.89±15.28	57.50±24.70	47.39±15.62
4	RNA synthesis (18)	19.92±9.28	28.57±13.80	22.53±9.54
5	DNA synthesis (7)	0	0	NaN
6	replication (40)	39.58±12.21	46.67±16.17	42.00±12.18
7	TCA cycle (11)	14.86±5.56	33.33±0	20.04±5.09
8	proteasome (35)	**79.87±9.50**	**90.00±9.47**	**84.20±7.29**
9	plastid (27)	**85.70±13.28**	**90.00±10.59**	**86.94±8.90**
10	merozoite Invasion (87)	**78.42±6.11**	**80.64±7.92**	**79.25±5.21**
11	Actin myosin motors (17)	37.46±16.86	37.93±16.34	35.65±13.33
12	Early ring transcripts (34)	**92.07±8.80**	**90.67±11.43**	**90.65±6.99**
13	Mitochondria (19)	13.97±5.64	27.14±9.94	18.02±6.52
14	Organellar translation (39)	34.24±10.70	41.82±11.60	37.03±9.68

We further tried to predict novel genes using the entire genome data of 6562 genes, using the benchmark data.

3.3 Bioinformatics data mining

In this initial proof of concept study on gene networks, we identified putative genes in the 14 selected classes that may represent different types of biological interactions. The Gene Ontology study supported the predictions. For examples:

(1) Glycolysis/TCA cycle and Nucleotide (DNA or RNA) synthesis processes exemplify metabolic networks which involve protein-metabolite interactions. For example, the presence of a cascade of co-expressed enzymes, including glucose-6-phosphate isomerase, glycerol-3-phosphate dehydrogenase, pyruvate kinase, lactate dehydrogenase, not only suggests that malaria parasite possesses conserved key components in carbohydrate metabolism, but also portrays the various co-factors and metabolites that are involved in the activity of each enzyme.

(2) Transcription, translation, and DNA replication machineries are complex networks that involve fine regulations of DNA (RNA)-protein and protein-protein interactions. For instance, Gene Ontology prediction suggested that, besides essential factors (e.g., initiation factors and elongation factors), other important components such as nascent polypeptide associated complex and peptide chain release factor may belong to translation machinery.

(3) Proteasome is a tightly-wrapped complex of threonine proteases and regulatory proteins that mediate protein-protein interactions in cell cycle control and stress response. In previous work [20], we predicted a number threonine proteases and ubiquitin hydrolases, sketching the core elements of malarial proteasome. A concerted regulation pattern revealed by this study is consistent with the postulation of an essential ATP-dependent ubiquitin-proteasome pathway, which was inferred from the results of inhibition assays [21].

In particular, we have explored the transcriptional machinery which is comprised by key enzymes and transcription factors which bind to the promoter elements, upstream elements and other proteins, and either facilitate or inhibit transcription.

Figure 3. Multiple alignment of the predicted Myb domain regions of the putative malaria gene PF10_0327, with 6 homologs. The arrows represent the two Myb regions. The boxes enclose the characteristic tryptophan residues. *P. berghei* and *P. yoelii* are rodent malaria parasites.

However, to date little is known about the transcriptional machinery in *P. falciparum*. Only 14 transcription factors were predicted by the P. *falciparum* genome annotation based on Gene Ontology. It seems implausible that this limited number of transcription factors represents the whole transcription factor repertoire, given the apparent need for extensive transcriptional control.

Using 23 putative genes in transcriptional machinery as a training set, SVM learning machine yielded 557 positive hits. These genes share similar profiles with peaks in ring and early-trophozoite stages, active stages for cascade transcriptional events.Our bioinformatics analysis suggested that the predicted genes belong to three categories:

(1) Genes involved in transcriptional process (Table 2): 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. Most interestingly, a putative transcription factor, PF10_0327, showed considerably high homology to the Myb and cdc5 proteins which both play multiple key roles in mitosis and cytokinesis. This prediction is reinforced by the observation of two Myb domains and the associated Tryptophan signature motifs, which are present in all known characterized Myb transcription factors (Figure 3).

(2) Genes involved in processes that are tightly associated with transcription. Several genes may be components of downstream processes such as pre-RNA processing after transcription.

Table 2: Putative genes predicted by SVM that may be involved in transcriptionary machinery.

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
n134_51	PF14_0612	putative zinc finger protein
f34582_1	MAL6P1.193	Zn-finger C2HC domain
m44300_14	PF13_0152	sir2 homologue
M33088_1	MAL13P1.213	transcription activator

(3) Genes that encode hypothetical proteins. By identifying co-expressed genes in developmental cycle, it also helps us to identify what could conceivably be network modules. Any network module could contain a range of proteins and regulatory elements [44]. The key components of these modules may have stringent functional constraint and hence are conserved across species [45, 46]. Subtracting these known from the modules, the remaining "hypothetical" in transcriptomic maps represent lineage-specific gaps in gene networks. The ability to assign a "hypothetical" gene to a specific network module opens an opportunity toward a tempo-specific functional characterization, because for a parasite with multiple hosts (human and mosquito) and a dynamic life cycle, "when and where" to initial wet-lab experiments is of critical importance. This network view should allow us to locate choke points in the parasite - potential vulnerabilities that

could result in new malarial control strategies.

4 Conclusion

This study complements the exciting model-based inference of genetic regulatory networks by enriching the list of network components. We have discovered novel network components from temporal expressional profiles using an integrated wavelet decomposition, SVM and bioinformatics data mining approach. These components could shed light on as yet unrecognized network interactions, and can serve as the starting point for model-based inference or/and functional characterization.

References:

[1] Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al., Genome sequence of the human malaria parasite *Plasmodium falciparum*, *Nature*, Vol. 419, No. 6906, 2002, pp. 498-511.

[2] Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, et al., Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*, *Nature*, Vol. 419, No. 6906, 2002, pp. 512-519.

[3] Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, et al., Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*., *Science*, Vol. 304, No. 5669, 2004, pp. 441-445.

[4] Kitano H, Systems biology: a brief overview, *Science*, Vol. 295, No. 5560, pp.1662-1664.

[5]Fraunholz MJ., Systems biology in malaria research. *Trends Parasitol.* Vol. 21, No. 9, 2005, pp. 393-395.

[6] Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, Berriman M, et al., A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses, *Science*, Vol. 307, No. 5706 , 2005, pp. 82-86.

[7] Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL., The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.*, Vol.1, No. 1, 2003, pp. E5.

[8] Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, et al., Discovery of gene function by expression profiling of the malaria parasite life cycle, *Science*, Vol. 301, No. 5639, 2003, pp1503-1508.

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

[10] Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D., Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proc Natl Acad Sci U S A.* Vol. 97, No. 1, 2000, pp262-267.

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

[12] Eisen MB, Spellman PT, Brown PO, Botstein D., Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci U S A.*, Vol. 95, No. 25, pp.14863-14868.

[13] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM., Systematic determination of genetic network architecture, *Nature Genetics*, Vol. 22, No.3, 1999, pp.281-285.

[14]Zhou X, Wang X, Dougherty ER, Construction of genomic networks using mutual information clustering and reversible-jump Markov-chain-Monte-Carlo predictor design, *Signal Processing*, vol. 83, 2003, pp. 745–761.

[15]Segal E, *Rich probabilistic models for genomic data*, Ph.D. thesis, Stanford University, 2004.

[16]Huang Y, Wang J, Wang Y, Zhang J, Bayesian inference of cell cycle regulatory networks, *IEEE Workshop on Genomic Signal Processing and Statistics*, 2005.

[17]Daubechie I, Ten Lectures on Wavelets, Society for Industrial and Applies Mathematics, 1992.

[18]Lu Y, Tian Q, Sanchez M, Wang Y, Hybrid PCA and LDA Analysis of Microarray Gene Expression Data, *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2005.

[19] The Gene Ontology Consortium, Gene Ontology: tool for the unification of biology, *Nature Genetics*, Vol. 25, 2000, pp. 25-29.

[20] Wu Y, Wang X, Liu X, Wang Y., Data-mining approaches reveal hidden families of proteases in the genome of malaria parasite, *Genome Research*. Vol. 13, No. 4, 2003, pp. 601-616.

[21]Gantt SM, Myung JM, Briones MR, Li WD, Corey EJ, Omura S, Nussenzweig V, Sinnis P., Proteasome inhibitors block development of *Plasmodium spp*. *Antimicrob Agents Chemother*. Vol. 42, No.10, 1998, pp:2731-2738.