Methods of Microarray Data Analysis V

Edited by Patrick McConnell, Simon Lin and Patrick Hurban

Since the inception of microarrays, studies in this field have drastically evolved with analysis methods needing to advance in-step. The CAMDA conference plays a role in this ever-changing discipline by providing a forum in which investigators can analyze the same datasets using different methods.

METHODS OF MICROARRAY DATA ANALYSIS V is the fifth book in this series, and focuses on the important issue of analyzing array data in a time series with correlating biological data. Previous books in this series focused on classification (Volume I), pattern recognition (Volume II), quality control issues (Volume III), and survival analysis (Volume IV).

In this volume, all investigators analyzed a single dataset on the lifecycle of the most deadly of malaria parasites, Plasmodium falciparum. The emphasis this year is on the application of novel and existing computational methodologies towards infectious disease. We highlight an introductory chapter by Raphael D. Isokpehi, a leading expert in the field of malaria. Ten of the papers presented at the conference are included, which range from the inference of genetic networks to the analysis of the spatial correlation of array data. This book is an excellent reference for academic and industrial researchers who want to keep abreast of the state-of-the-art in microarray data analysis.

Patrick McConnell is a researcher in the Duke Bioinformatics Group in the Duke Comprehensive Cancer Center.

Simon M. Lin is a faculty member in the Robert H. Lurie Comprehensive Cancer Center and associate director of Bioinformatics at Northwestern University.

Patrick Hurban is director of Investigational Genomics at Icoria, Inc., a Clinical Data company.

ISBN 0-387-34568-X

9 780387 345680

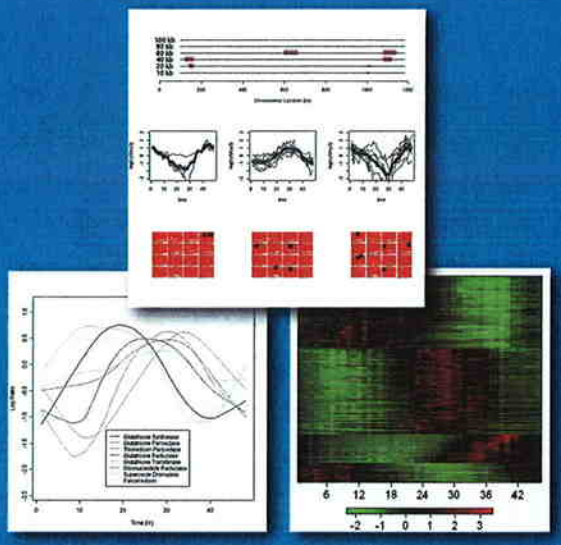
springer.com

Methods of Microarray Data Analysis V

McConnell / Lin / Hurban

Methods of Microarray Data Analysis V

Edited by
Patrick McConnell, Simon Lin and Patrick Hurban



Springer

Library of Congress Control Number: 2005297773

ISBN:10: 0-387-34568-X e-ISBN-10: 0-387-34569-8
ISBN-13: 978-0-387-34568-0 e-ISBN-13: 978-0-387-34569-7

Printed in acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Contents

Silvestrini, F., Bozdech, Z., Lanfrancotti, A., Di, G.E., Bultrini, E., Picci, L., Derisi, J.L., Pizzi, E., and Alano, P. (2005), Genome-wide identification of genes upregulated at the onset of gametocytogenesis in *Plasmodium falciparum*, *Mol. Biochem. Parasitol.*, **143**, 100–110.

Snow, R.W., Guerra, C.A., Noor, A.M., Myint, H.Y., and Hay, S.I. (2005), The global distribution of clinical episodes of *Plasmodium falciparum* malaria, *Nature*, **434**, 214–217.

Talman, A.M., Domarle, O., McKenzie, F.E., Ariey, F., and Robert, V. (2004), Gametocytogenesis: The puberty of *Plasmodium falciparum*, *Malar. J.* **3** (24), 24.

Waller, R.F., and McFadden, G.I. (2005), The apicoplast: A review of the derived plastid of apicomplexan parasites, *Curr. Issues Mol. Biol.*, **7**, 57–79.

Wang, R., Revalo-Herrera, M., Gardner, M.J., Bonelo, A., Carlton, J.M., Gomez, A., Vera, O., Soto, L., Vergara, J., Bidwell, S.L., Domingo, A., Fraser, C.M., and Herrera, S. (2005), Immune responses to *Plasmodium vivax* pre-erythrocytic stage antigens in naturally exposed Duffy-negative humans: A potential model for identification of liver-stage antigens, *Eur. J. Immunol.*, **35**, 1859–1868.

Webster, D. (2001), Malaria kills one child every 30 seconds, *J. Public. Health Policy*, **22**, 23–33.

WHO (2005), World Malaria Report 2005, World Health Organization, Geneva.

Yeh, I., Hanekamp, T., Tsoka, S., Karp, P.D., and Altman, R.B. (2004), Computational analysis of *Plasmodium falciparum* metabolism: Organizing genomic information to facilitate drug discovery, *Genome Res.*, **14**, 917–924.

Young, J.A., Fivelman, Q.L., Blair, P.L., De V, L., Le Roch, K.G., Zhou, Y., Carucci, D.J., Baker, D.A., and Winzeler, E.A. (2005), The *Plasmodium falciparum* sexual development transcriptome: A microarray analysis using ontology-based pattern identification, *Mol. Biochem. Parasitol.*, **143**, 67–79.

Chapter 2

Constructing Probabilistic Genetic Networks of *Plasmodium falciparum* from Dynamical Expression Signals of the Intraerythrocytic Development Cycle

Junior Barrera[a], Roberto M. Cesar Jr.[a], David C. Martins Jr.[a], Ricardo Z.N. Vêncio[a], Emilio F. Merino[b], Márcio M. Yamamoto[b], Florencia G. Leonardi[a], Carlos A. de B. Pereira[a] and Hernando A. del Portillo[b]

[a]*Instituto de Matemáticas e Estatística e BIOINFO-USP, Núcleo de Pesquisas em Bioinformática, Universidade de São Paulo, R. do Matão 1010, São Paulo, SP 05508-900, Brazil*

[b]*Instituto De Ciências Biomédicas, Departamento de Parasitologia, Universidade de São Paulo, Ave. Lineu Prestes 1374, São Paulo, SP 05508-900, Brazil*

Abstract

The completion of the genome sequence of *Plasmodium falciparum* revealed that close to 60% of the annotated genome corresponds to hypothetical proteins and that many genes, whose metabolic pathways or biological products are known, have not been predicted from sequence similarity searches. Recently, using global gene expression of the asexual blood stages of *P. falciparum* at 1 h resolution scale and Discrete Fourier Transform based techniques, it has been demonstrated that many genes are regulated in a single periodic manner during the asexual blood stages. Moreover, by ordering the genes according to the phase of expression, a new list of targets for vaccine and drug development was generated. In the present paper, genes are annotated under a different perspective: a list of functional properties is attributed to networks of genes representing subsystems of the *P. falciparum* regulatory expression system. The model developed to represent genetic networks, called Probabilistic Genetic Network (PGN), is a Markov chain with some additional properties. This model mimics the properties of a gene as a non-linear stochastic gate and the systems are built by coupling of these gates. Moreover, a tool that integrates mining of dynamical expression signals by PGN design techniques, different databases and biological knowledge, was developed. The applicability of this tool for discovering gene networks of the malaria expression regulation system has been validated using the glycolytic pathway as a "gold-standard", as well as by creating an apicoplast PGN network. Presently, we are tentatively improving the network design technique before trying to validate results from the apicoplast PGN network through reverse genetics approaches.

Keywords: malaria, annotation tool, probabilistic genetic networks, dynamical system, Markov chain, mutual information, gene expression, microarrays

1. INTRODUCTION

Malaria remains the most devastating parasitic disease worldwide, and every year is responsible for 300–500 million clinical cases and 1–2 million deaths, mostly in children below 5 years old [http://www.who.int/tdr/diseases/malaria/default.htm]. Furthermore, the appearance of drug-resistant parasite strains to most antimalarial drugs and of insecticide-resistant *Anopheles* mosquitoes, in addition to the global warming, all have exacerbated this public health situation.

The advent of genomics into malarial research is significantly accelerating the discovery of control strategies. Indeed, the first draft of the complete genome sequence of *Plasmodium falciparum*, the most deadly human malaria parasite, was released only three years ago [9], but it has substantially modified the way of thinking for the development of new vaccines, drugs and alternatives of control strategies. Moreover, it has allowed the initiative for global scale studies on the transcriptome [1,4,10,12,16], proteome [7,10,14,15] and metabolome [23] of the parasite in different developmental stages.

Recent experimental evidence indicates that malaria parasites present unique mechanisms for control of gene expression: data from SAGE analysis has demonstrated that approximately 17% of abundant tags correspond to anti-sense transcripts of annotated genes [17], that suggests that these anti-sense transcripts might be involved in post-transcriptional regulation; reverse genetics approaches have shown that introns co-regulate expression of variant genes [2]; although promoters seem to be bi-partite, it is postulated that there must be unique sets of malarial transcription factors due to the high AT-content of intergenic regions and absence of recognized regulatory transcription factors [3,13].

Progressing the research effort, dynamical global gene expression measures of the intraerythrocytic developmental cycle (IDC) of the parasite at 1 h-scale resolution were recently reported [1]. Moreover, using Discrete Fourier Transform (DFT) based techniques, researchers verified that, during this life stage, the parasite seems to follow a rigid clockwise program where genes with common functions are transcribed at similar times. This study recognized 73% of the quality controlled (QC) dataset available for the CAMDA contest (http://www.camda.duke.edu/camda04/datasets/). The QC dataset comprises 3719 elements with relative expression signals with almost sinusoidal shape in the logarithmic scale or, equivalently, pulse like shape in the original relative expression scale. By ordering these signals by phase, they constructed a wave

of signal propagation and ordered genes. Analysis of ordered genes throughout the asexual blood stages provided a comprehensive and biologically meaningful list of genes with putatively similar functions [1]. This analysis, however, did not include the elements that did not have almost sinusoidal shape and which, however, represented 27% of the QC dataset (i.e., 1361 elements).

In this paper, a list of functional properties is attributed to networks instead of individual genes. To do so, a tool was created that integrates mining procedures of dynamical expression signals and conventional databases (i.e., genome, proteome, metabolome, and clinical data).

This annotation approach may be applied to all spotted oligonucleotides of the QC set, despite the shape of their dynamical signals being sinusoidal or not. Subsystems of the malaria expression regulation system are modeled as probabilistic genetic networks (i.e., a stochastic process that is a specialization of a Markov chain) [20]. These networks are designed from the observed dynamical signals. The designed subsystems are annotated using conventional public databases and biological knowledge. The subsystems to be designed are defined from seed genes of particular biological interest, i.e. the subsystems are composed by genes that predict or are predicted by seed genes [11]. For example, some genes analyzed by the DFT approach were used as seeds to discover other non-sinusoidal genes associated to the same phase of the parasite life cycle.

Following this Introduction, Section 2 presents the concept of probabilistic genetic network (PGN). Section 3 describes the technique used for designing a PGN. Section 4 describes the developed software tools. Section 5 gives results of the application of the design techniques to simulated PGNs and presents preliminary biological results obtained by applying the proposed technique. Finally, the results and future steps of this research are discussed in the Concluding Remarks.

2. PROBABILISTIC GENETIC NETWORKS

The life of an organism depends on many metabolic pathways that are regulated by gene expression networks. The mechanism of pathways regulation involves a complex system with many forward and feedback signals. These signals are RNA, produced by gene expression, and protein complexes, produced by interaction of proteins built by translation of mRNA. Protein complexes act as feedback signals that control gene transcription. Forward signals, in the form of enzymes, act as control metabolic pathways. In such networks, the expression of each gene depends both on its own expression and on the expression levels of other genes at previous time instants. This complex network of interactions can thus be modeled by a dynamical system.

Finite dynamical systems, discrete in time and finite in range, can model the behavior of gene expression networks. In such model, we represent each transcript by a variable that takes the expression value of that transcript. All these variables, taken collectively, are the components of a vector called the *state of the system*. Each component (i.e. transcript) of the state vector has an associated function that calculates its next value (i.e. expression value) from the state at previous time instants. These functions are the components of a function vector, called *transition function*, which defines the transition from one state to the next and represents the gene regulation mechanisms. In order to formalize these ideas, we introduce some definitions and notation. Let R be the range of all state components. For example, $R = \{0, 1\}$ in binary systems, and $R = \{-1, 0, 1\}$ in three levels systems. The transition function ϕ, for a gene network of n genes, is a function from R^n to R^n. A finite dynamical system is given by

$$x[t+1] = \phi\big(x[t]\big),$$

where $x[t] \in R^n$, for every $t \geqslant 0$. A component of $x[t]$ is a value $x_i[t] \in R$.

Systems defined as above are *time translation invariant*, i.e. the transition function is the same for all discrete time t. When ϕ is a stochastic function (i.e. for each state $x[t]$, the next state $\phi(x[t])$ is a realization of a random vector), the dynamical system is a stochastic process.

In this paper, we represent gene expression networks by stochastic processes. The stochastic transition function is a particular family of Markov chains called probabilistic genetic network (PGN).

Consider a sequence of random vectors $X_0, X_1, X_2, \ldots$ assuming values in R^n and its realizations denoted, respectively, $x[0], x[1], x[2], \ldots$. A sequence of random states $(X_t)_{t=0}^{\infty}$ is called a Markov chain if, for every $t \geqslant 1$,

$$P\big(X_t = x[t] | X_0 = x[0], \ldots, X_{t-1} = x[t-1]\big) = P\big(X_t = x[t] | X_{t-1} = x[t-1]\big).$$

The significance of a Markov chain lies in the fact that the conditional probability of the future event, given the past history, depends only upon the immediate past and not upon the remote past.

A Markov chain is characterized by a transition matrix $\pi_{Y|X}$ of conditional probabilities between states, whose elements are denoted $p_{y|x}$, and an initial condition random vector of states π_0. The stochastic transition function ϕ at the time t is given by

$$\phi\big(x[t]\big) = y,$$

for every $t \geqslant 1$, where y is a realization of a random vector with distribution $p_{\cdot|x[t]}$.

A *Probabilistic Genetic Network* (PGN) is a Markov chain $(\pi_{Y|X}, \pi_0)$ such that

(i) $\pi_{Y|X}$ is homogeneous, i.e. $p_{y|x}$ is not a function of t.
(ii) $p_{y|x} > 0$, for every pair of states $x, y \in R^n$.
(iii) $\pi_{Y|X}$ is conditionally independent, i.e. for every pair of states $x, y \in R^n$,

$$p_{y|x} = \prod_{i=1}^{n} p(y_i|x).$$

(iv) $\pi_{Y|X}$ is almost deterministic, i.e. for every state $x \in R^n$, there exists a single state, $y \in R^n$ such that $p_{y|x} \approx 1$.
(v) For every gene j there exists a vector a^j of integer numbers such that for every $x, z \in R^n$ and $y_j \in R$,

$$\text{if} \quad \sum_{i=1}^{n} a_i^j x_i = \sum_{i=1}^{n} a_i^j z_i \quad \text{then} \quad p(y_j|x) = p(y_j|z).$$

These axioms imply that each gene is characterized by a vector of coefficients a and a vector stochastic function g_j from Z, the set of integers numbers, to R. If a_i^j is positive then the target gene j is *excited* by gene i. If a_i^j is negative then it is *inhibited* by gene i. If a_i^j is 0, then it is *not affected* by gene i. We say that gene j is *predicted* by the gene i when a_i^j is different of 0. The component j of the stochastic transition function ϕ, denoted ϕ_j, is built by the composition of g_j with the linear combination of a^j and the previous state $x[t]$, i.e. for every $t \geqslant 1$,

$$\phi_j\big(x[t]\big) = g_j\left(\sum_{i=1}^{n} a_i^j x_i[t]\right),$$

where $g_j(\sum_{i=1}^{n} a_i^j x_i[t])$ is a realization of a random variable in R, with distribution $p(\cdot|\sum_{i=1}^{n} a_i^j x_i[t])$.

The axioms that define the PGN model are inspired in biological phenomena or mandatory simplifications due to the usual lack of data for the model estimation. The main hypothesis adopted is to choose a discrete model. This is justified because transcription and translation are discrete phenomena. The

levels of quantization are chosen according to the available data for model parameters estimation.

Axiom (i) is a constraint just to simplify the estimation problem, but it could be generalized easily. Axiom (ii) imposes that all states are reachable, that is, noise may lead the system to any state. It is a quite general model that reflects our lack of knowledge about the kind of noise that may affect the system. Axiom (iii) means that the expression of a gene at a given time instant t does not depend on the expression of other genes at t. This happens when the time step of the model is less than the time spent for transcription-translation. Axiom (iv) means that the system has a main structural dynamics that is subject to small noise. This is what happens in practically all known engineering systems designed by man. Axiom (v) means that genes act as a non linear gate triggered by a balance between inhibitory and excitatory inputs, analogous to neurons.

It is important to recall that axiom (iii) might no be verified due to the 1h time resolution limitation of the available experimental data. However, this axiom was adopted in our model to allow statistical tractability. Of importance, using this axiom, we were able to generate biologically meaningful results (see below).

3. DESIGN OF PGNS

The goal of this research is to estimate a PGN representing a subsystem of the malaria parasite gene expression network from dynamical microarray relative expression measures and biological knowledge. In the following the procedure adopted for PGN estimation is described.

The *entropy* $H(X)$ of a random variable X is a measure of its distribution $\{p_i\}$, given by

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i.$$

The entropy has some remarkable properties: (i) all the distributions formed by permutations of p_i have the same entropy; (ii) concentrating the probability mass of a distribution implies in decreasing its entropy. As a corollary of property (ii), the uniform distribution presents maximum entropy and those with minimum entropy have the total probability mass concentrated in one point.

The *mutual information* [5] between two random variables X and Y is the measure defined by

$$I(X, Y) = H(Y) - H(Y|X).$$

It measures the probability mass concentration of $P(Y)$ in $P(Y|X)$ by the observation of X. The expectation $E[I(X, Y)]$ of $I(X, Y)$ is given by

$$E\big[I(X, Y)\big] = H(Y) - E\big[H(Y|X)\big].$$

When $E[I(X, Y)] = 0$, X and Y may be independent variables and the condition $P(Y) = P(Y|X)$ should be tested. In case this condition is true, then X and Y are independent, otherwise, they have dependence.

The expectation of the mutual information is used to estimate the PGN. The random variable Y will be the gene value $y_i[t+1]$ to be predicted and the given random variable X will be the vector of genes $x[t]$ weighted by an integer vector$a$, associated to gene y_i. For each vector a, with $a_i \in \{-1, 0, +1\}$ and at most three values different from 0, the mean mutual information is estimated. The first vectors a, that have larger mutual information, are selected. These vectors indicate the connection between genes and the kind of connection: excitatory or inhibitory.

4. DEVELOPED SOFTWARE TOOLS

The designed software system estimates gene networks from dynamical expression measures and represents them as graphs linked to malaria databases. Firstly, the system receives the raw fluorescence intensity measures as input and applies a quality control procedure that generates a new dataset. Then, the signals of this dataset are normalized and quantized into three expression levels $\{-1, 0, +1\}$.

Some target genes together with the quantized signals are provided to the main module of the system, which is responsible for computing the best predictors set for each target (based on the PGN design techniques described in the last section).

A user-friendly graphical interface was implemented to facilitate the biological interpretation of the results. The table of predictors, the file of functional groups annotated by Bozdech et al. [1] and the Overview dataset (http://www.camda.duke.edu/camda04/datasets) are organized and given as input for the GraphViz (a package to visualize graphs, http://www.research.att.com/sw/tools/graphviz). A color code was assigned to each node of the network (i.e. oligo) according to the functional biological categories defined in [1]: transcriptional machinery (pink), cytoplasmic translation machinery (blue), glycolitic pathways (yellow), etc. (see Figure 2). Besides, the node shape indicates if the oligonucleotide is present in the Overview set or not: a square indicates that it is present and a circle that it is not. Each node has a link to a page with pointers to three public databases: PlasmoDB (http://plasmodb.org),

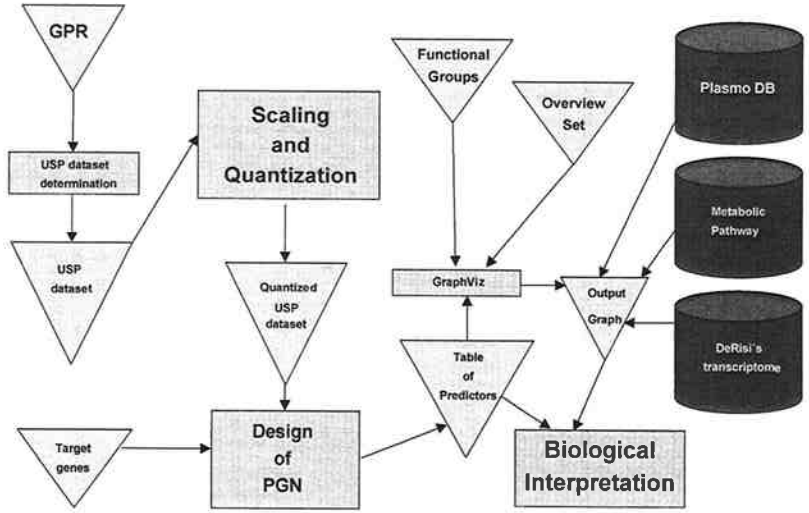


Figure 1. Simplified scheme of the analysis pipeline. Modules (squares) datafiles (triangles) databases (cylinders) and dataflow (arrows) of the data analysis pipeline.

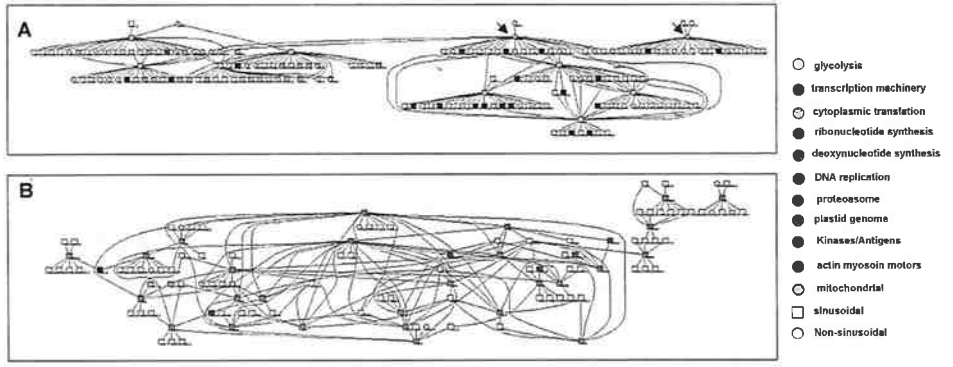


Figure 2. Glycolytic and Apicoplast Probabilistic Genetic Networks. Glycolytic (**A**) and Apicoplast (**B**) PGN networks were obtained by using those genes listed under functional group for glycolysis and plastid genome at the malaria IDC database (http://malaria.ucsf.edu/) as targets. Next, the five best tuples (individuals) of predictors were computed according to the mutual information criterion and a graph generated as described in Section 5.1. To facilitate visualization of the PGN networks, an arbitrary color-code was created to represent functional groups. Moreover, squares represent oligonucleotides with sinusoidal expression signals whereas circles represent oligonucleotides with non-sinusoidal expression signals. Glycolytic and Apicoplast networks can be visualized at http://www.vision.ime.usp.br/CAMDA2004/glycolysis.html and http://www.vision.ime.usp.br/CAMDA2004/apicoplast.html, respectively.

Metabolic Pathways (http://biocyc.org/PFA/) and DeRisi's transcriptome database (http://malaria.uscf.edu/). The output including the graph and links to public databases is fully generated in HTML.

Thus, this software allows easy access to different information of each target gene and can help in the annotation of hypothetical proteins and null elements. Figure 1 represents a scheme of the data analysis pipeline used in this study.

5. EXPERIMENTAL RESULTS

5.1. Simulations

For validating the proposed PGN estimation technique, artificial networks that satisfy the PGN definition were created, simulated and estimated. These simulated networks have 12 genes that may be predicted from one to five genes or may even be independent. All network genes are ternary (values in $\{-1, 0, +1\}$) and $p(y_i|X)$ has at least 80% of concentration mass. The simulations were just 48 iterations long (i.e. the number of iterations present at an 1 h-scale resolution observation of the asexual blood stages of *P. falciparum*). For each target gene, the five best tuples (individual, pairs, triples) of predictors were computed according to the mutual information criterion. The *quality of a predictor* g was defined as the addition of the mutual information of all tuples of predictors in which g appears. Finally, the predictors were ordered by their quality. In the performed experiments, the genes with greater quality were almost always exactly the predictors for the target gene. Some of these experiments can be found at the following site http://www.vision.ime.usp.br/CAMDA2004/simulations/.

5.2. Pre-processing

We performed standard pre-processing procedures in the contest dataset such as filtering low-intensity unreliable spots and dye bias normalization. Moreover, we checked the normalization procedure described by Bozdech et al. [1] and found that they used an overall global normalization factor to normalize the expression ratios. There are known concerns in using global normalization procedures since it could represent clear systematic non-linear dependence between expression ratio and fluorescence intensities [18]. However, we verified that non-linear dependences were negligible in the complete dataset available for CAMDA.

Bozdech et al. [1] excluded the low hybridization intensity signals since they received the same treatment as blobs or blotches. However, important biological information may be hidden in genes that are not expressed during some part of the intraerythrocytic developmental cycle of malaria parasites. We constructed a different dataset from the original output of GenePix. Original flags

for blobs or blotches were kept, whereas non-detectable expression values of low intensity signals were set to zero. We classified a spot as non-detectable if the mean intensity measurement in Cy3 or Cy5 is below to some local threshold value. This threshold is calculated from the distribution of pixel intensities of the background surrounding the spot. The 90% quantile of the local background distribution was used to define the intensity threshold. Spot's mean intensities below to this threshold were truncated to 0. This simple rule can naturally exclude unreliable signals, since the hybridization log-ratio $\log 2(0/0)$ is not defined. However, this rule preserves the potentially relevant situations when a signal is transcriptionally inactive only in a fraction of the time-course, since the expression becomes $\log 2(0/\text{reference}) = -\infty$. Although this is not a numerical ratio, the result can be incorporated in our Markovian approach because of the quantization step. As a result of this pre-processing step the USP-dataset used for the contest contained 6532 oligos, including 1361 oligos with not almost sinusoidal expression, as opposed to the 3719 oligos used in the overview dataset used by Bozdech et al. [1] to generate the phaseogram of the IDC malaria cycle.

5.3. Signal Normalization and Quantization

In order to validate the proposed methodology, the well known glycolytic pathway was studied. Before applying the predictor estimation techniques the signal was normalized and quantized. The signals were normalized by the *normal transformation* η given by, for every signal $g(t)$, $\eta[g(t)] = \frac{g(t)-E[g(t)]}{\sigma[g(t)]}$, where $E[g(t)]$ and $\sigma[g(t)]$ are, respectively, the expectation and standard deviation of $g(t)$.

The normal transformation has two important properties: (i) $E[\eta[g(t)]] = 0$ and $\sigma[\eta[g(t)]] = 1$, for every random variable $g(t)$; (ii) $\eta[g(t)] = \lambda\eta[g(t)]$, for every real number λ. The quantization of a gene at a given instant is a mapping from the continuous expression log-ratio into three qualitative expression levels $\{-1, 0, +1\}$ (i.e. down, null and up regulated in relation to the reference, respectively). The quantization of a gene signal g is performed by a *threshold mapping* given by

$$g'(t) = \begin{cases} +1 & \text{if } g(t) \geqslant h, \\ 0 & \text{if } l \leqslant g(t) \leqslant h, \\ -1 & \text{if } g(t) \leqslant l, \end{cases}$$

for every $t \geqslant 0$, where

$$l = \frac{\sum_{g(t)<0} g(t)}{|\{g(t):\ g(t) < 0\}|} \quad \text{and} \quad h = \frac{\sum_{g(t)>0} g(t)}{|\{g(t):\ g(t) > 0\}|}.$$

Normalization and quantization have the effect of creating equivalence classes between signals, thus decreasing estimation errors due to lack of data.

5.4. Glycolytic PGN Network

During the asexual blood stages, malaria parasites rely entirely on glycolysis for its ATP production [21]. Thus, we chose target genes that code for all the 10 enzymes pertaining to the glycolytic pathway (hexokinase, phosphohexose isomerase, phosphofructokinase1, aldolase, triose phosphate isomerase, glyceraldehide 3 phosphate dehydrogenase, phosphoglycerate kinase, phosphoglycerated mutase, enolase, and pyruvate kinase) to test our model. Significantly, an interconnected glycolytic PGN network was generated by using the first five genes with the lowest entropic values associated to each glycolytic enzyme (Figure 2). Moreover, analysis of 40 best predictors for each glycolytic target (289 distinct oligos in total) revealed that most of them (96%) corresponded to hypothetical proteins genes related to transcription, translation, DNA and RNA synthesis, actin myosin motors and kinases (http://www.vision.ime.usp.br/CAMDA2004/Table1S.html). Remaining genes encoded surface antigens and thus, *a priori*, can be considered false-positives. Worth mentioning, similar results were obtained from a list of 400 genes expressed in-phase with glycolysis obtained from data of Bozdech et al. [1] (not shown). As expected, no genes of the TCA cycle were found in the glycolytic PGN network further corroborating the lack of a functional TCA cycle during the asexual blood stages of malaria parasites [21]. Of relevance, several oligonucleotides not included in the overview dataset due to low hybridization intensity or non-sinusoidal signals, were included in the PGN network (Figure 2). Of relevance, two oligonucleotides (opff72413 and m11919_1) corresponding to two glycolytic enzymes, hexokinase and aldolase, respectively, excluded from the phaseogram of Bozdech et al. [1], were included in the glycolitic PGN network. Together, this data demonstrates the value of the PGN model in generating a biologically meaningful glycolytic network that includes genes not included by the Fourier approach [1].

Next, we attempted to create an apicoplast PGN network. Enzymes from this organelle are becoming new targets for malaria since there is no homologous organelle in the human host [19,22]. Of relevance, two different computational algorithms have been developed to predict apicoplast proteins. In the first one, a genome-wide scan of *P. falciparum* revealed over 550 nuclear genes that encoded a consensus bi-partite peptide signal sequence [8]. In the second one, genes expressed in-phase with the plastid genome and containing the bi-partite peptide signal sequences narrowed the list of apicoplast nuclear-encoded genes from over 550 to 156 [1]. In order to apply our algorithm, oligonucleotides representing each of the 20 putative apicoplast genome-encoded proteins listed

from DeRisi's laboratory (http://malaria.ucsf.edu/) were fed to our program and an apicoplast PGN network was generated (Figure 2). Analysis of the results clearly indicated that our method is capable of interconnecting genes that have been experimentally demonstrated to be part of the apicoplast (acyl-carrier protein and ribosomal protein S9), whereas many other genes lack predicted bi-partite peptide signal sequences. These results are difficult to reconcile with our present knowledge of the predicted malaria apicoplast proteome. Reverse genetics approaches similar to the ones used to define the importance of the bi-partite peptide signal sequences [8] can now be envisioned to validate some of these genes. Alternatively, our model describes genes not only related to the apicoplast proteome but genes whose expression is essential to create such network.

As our program creates PGN networks, a negative control was idealized to further validate the biological value of our findings. Thus, eight genes, four from glycolysis and four from the apicoplast organelle, were chosen randomly and used together as seed genes to create PGN networks based on single-gene and two-gene predictions. The results clearly demonstrated that the glycolysis and apicoplast PGN networks based on single-gene predictions were not interconnected (http://www.vision.ime.usp.br/CAMDA2004/ga.html). Based on two-gene predictions, with the exception of two genes from the glycolytic PGN network that inter-connected with the apicoplast PGN network, remaining genes were not connected (http://www.vision.ime.usp.br/CAMDA2004/ga2.html). It is important to recall that two-gene predictions are based on 21,330,246 calculations further reinforcing the value of these results. Together, this data demonstrates the value of the PGN model in generating biologically meaningful networks and which include genes not included by the Fourier approach [1].

An ideal PGN network will include interconnectivity networks based on interactions of several genes. Unfortunately, the "limited" amount of data presently available from the IDC transcriptome of *P. falciparum*, precludes such analyses without introducing a large degree of error. Regardless, this data demonstrates that the PGN model and program presented here are capable of constructing biologically meaningful networks of malaria from dynamical expression signals of the asexual blood stages and that it can be used as a complementary computational approach to Fourier analysis by including genes that are not periodically expressed.

6. CONCLUDING REMARKS

In order to advance our knowledge on the biology of *P. falciparum*, we have designed PGNs from dynamical expression signals of the asexual blood stages

reported by Bozdech et al. [1]. Unlike their DFT approach, PGN design allowed us to use all the elements available in the dataset. Significantly, this technique was applied to target genes that code for enzymes of the glycolytic pathway and a biologically meaningful glycolytic network was obtained. Next, we applied this algorithm to construct an apicoplast PGN network and although "signature" apicoplast genes were found, many other genes lack the consensus bipartite peptide signal sequence.

These results were obtained without considering the equivalence between linear combinations of inputs, what should improve the results, since the estimation errors will diminish and the hypothesis is quite consistent with observed gene dynamics. Besides, this model will permit to distinguish between inhibitory and excitatory signals. Although the normal transformation creates equivalence classes that diminishes the estimation errors, it amplifies noise in housekeeping genes that have almost constant expression signals. One way of circumventing this problem is to detect and exclude the housekeeping genes of the regulatory systems study before signal quantization.

The next steps of this research include mainly improving the network design technique and validation through reverse genetics approaches of some of the genes previously unpredicted by other algorithms as being part of the apicoplast. If validated, the PGN approach could thus be used to annotate genes not considered by the DFT approach and to accelerate the discovery of new targets against malaria.

ACKNOWLEDGEMENTS

Our thanks to Professor Bianca Zingales (IQ-USP) for first catalysing the interaction between JB and HAP, and to TDR-WHO for consolidating the interaction between the Departments of Parasitology and Computer Science through a Grant on Bioinformatics and Tropical Diseases (http://malariadb.ime.usp.br/courses.html). The authors also thank FAPESP (01/09401-0, 98/12765-8, 04/03967-0, 02/04698-8), CNPq and CAPES by their continuous support.

REFERENCES

[1] Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J., and DeRisi, J.L., The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*, *PLoS Biol.*, **1** (2003), 5.

[2] Calderwood, M.S., Gannoun-Zaki, L., Wellems, T.E., and Deitsch, K.W., *Plasmodium falciparum* var genes are regulated by two regions with separate promoters, one upstream of the coding region and a second within the intron, *J. Biol. Chem.*, **278**(36) (2003), 34125–34132.

[3] Coulson, R.M., Hall, N., and Ouzounis, C.A., Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*, *Genome Res.*, **14** (2004), 1548–1554.

[4] Daily, J.P., Le Roch, K.G., Sarr, O., Fang, X., Zhou, Y., Ndir, O., Mboup, S., Sultan, A., Winzeler, E.A., and Wirth, D.F., In vivo transcriptional profiling of *Plasmodium falciparum*, *Malar J.*, **3** (2004), 30.

[5] DeGroot, M.H., Uncertainty, information and sequential experiments, *Ann. Math. Statist.*, **3** (1962), 404–419.

[6] Dougherty, E.R., Bittner, M.L., Chen, Y., Kim, S., Sivakumar, K., Barrera, J., Meltzer, P., and Trent, J.M., In: *Proceedings of Nonlinear Filters in Genomic Control. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing* (Antalia, Turkey, 1999), pp. 10–15.

[7] Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D., Moch, J.K., Muster, N., Sacci, J.B., Tabb, D.L., Witney, A.A., Wolters, D., Wu, Y., Gardner, M.J., Holder, A.A., Sinden, R.E., Yates, J.R., and Carucci, D.J., A proteomic view of the *Plasmodium falciparum* life cycle, *Nature*, **419** (2002), 520–526.

[8] Foth, B.J., Ralph, S.A., Tonkin, C.J., Struck, N.S., Fraunholz, M., Roos, D.S., Cowman, A.F., and McFadden, G.I., Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*, *Science*, **299** (2003), 705–708.

[9] Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., and Barrell, B., Genome sequence of the human malaria parasite *Plasmodium falciparum*, *Nature*, **419** (2002), 498–511.

[10] Hall, N., Karras, M., Raine, J.D., Carlton, J.M., Kooij, T.W., Berriman, M., Florens, L., Janssen, C.S., Pain, A., Christophides, G.K., James, K., Rutherford, K., Harris, B., Harris, D., Churcher, C., Quail, M.A., Ormond, D., Doggett, J., Trueman, H.E., Mendoza, J., Bidwell, S.L., Rajandream, M.A., Carucci, D.J., Yates III, J.R., Kafatos, F.C., Janse, C.J., Barrell, B., Turner, C.M., Waters, A.P., and Sinden, R.E., A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic and proteomic analyses, *Science*, **307** (2005), 82–86.

[11] Hashimoto, R.F., Kim, S., Shmulevich, I., Zhang, W., Bittner, M.L., and Dougherty, E.R., Growing genetic regulatory networks from seed genes, *Bioinformatics*, **20** (2004), 1241–1247.

[12] Hayward, R.E., Derisi, J.L., Alfadhli, S., Kaslow, D.C., Brown, P.O., and Rathod, P.K., Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria, *Mol. Microbiol.*, **35** (2000), 6–14.

[13] Horrocks, P., Dechering, K., and Lanzer, M., Control of gene expression in *Plasmodium falciparum*, *Mol. Biochem. Parasitol.*, **95** (1998), 171–181.

[14] Lasonder, E., Ishihama, Y., Andersen, J.S., Vermunt, A.M., Pain, A., Sauerwein, R.W., Eling, W.M., Hall, N., Waters, A.P., Stunnenberg, H.G., and Mann, M., Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry, *Nature*, **419** (2002), 537–542.

[15] Le Roch, K.G., Johnson, J.R., Florens, L., Zhou, Y., Santrosyan, A., Grainger, M., Yan, S.F., Williamson, K.C., Holder, A.A., Carucci, D.J., Yates, J.R., III, and Winzeler, E.A., Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle, *Genome Res.*, **14** (2004), 2308–2318.

[16] Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holder, A.A., Batalov, S., Carucci, D.J., and Winzeler, E.A., Discovery of gene function by expression profiling of the malaria parasite life cycle, *Science*, **301** (2003), 1503–1508.
[17] Patankar, S., Munasinghe, A., Shoaibi, A., Cummings, L.M., and Wirth, D.F., Serial analysis of gene expression in *Plasmodium falciparum* reveals the global expression profile of erythrocytic stages and the presence of anti-sense transcripts in the malarial parasite, *Mol. Biol. Cell.*, **12** (2001), 3114–3125.
[18] Quackenbush, J., Microarray data normalization and transformation, *Nat. Genetics*, **32** (2002), 496–501.
[19] Ralph, S.A., Van Dooren, G.G., Waller, R.F., Crawford, M.J., Fraunholz, M.J., Foth, B.J., Tonkin, C.J., Roos, D.S., and McFadden, G.I., Metabolic maps and functions of the *Plasmodium falciparum* apicoplast, *Nat. Rev. Microbiol.*, **2** (2004), 203–216.
[20] Shimulevich, I., Dougherty, E.R., Kim, S., and Zhang, W., Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks, *Bioinformatics*, **18**(2) (2002), 261–274.
[21] Sherman, I.W., Metabolism and surface transport of parasitized erythrocytes in malaria, *Ciba Found Symp.*, **94** (1983), 206–221.
[22] Wilson, R.J.M. (Iain), Progress with parasite plastids, *J. Mol. Biol.*, **319** (2002), 257–274.
[23] Yeh, I., Hanekamp, T., Tsoka, S., Karp, P.D., and Altman, R.B., Computational analysis of *Plasmodium falciparum* metabolism: Organizing genomic information to facilitate drug discovery, *Genome Res.*, **14** (2004), 917–924.

Chapter 3

Simple Methods for Peak and Valley Detection in Time Series Microarray Data

A. Sboner[a,b], A. Romanel[b], A. Malossini[b], F. Ciocchetta[a,b], F. Demichelis[a,b], I. Azzini[a], E. Blanzieri[b] and R. Dell'Anna[a]

[a]*Bioinformatics Group, SRA Division, ITC-irst, Via Sommarive 18, I-38050 Povo (TN), Italy*
[b]*Department of Information and Communication Technology, University of Trento, Via Sommarive 14, I-38050 Povo (TN), Italy*

Abstract Given a set of gene expression time series obtained by a microarray experiment, this work proposes a novel quality control procedure that exploits six analytical methods, each of which allows for the identification in an automated way of genes that have expression spikes within narrow time-windows and over a chosen amplitude threshold. The output of these methods, suitably combined in an automated way, provides an exhaustive list of genes and time points in which abrupt variations have been detected. The quality control on these genes is then performed by a biologist, who classifies the spikes either as biologically relevant or as artifacts. In the latter case, spikes must be eliminated by a smoothing procedure. In this chapter, we first describe the six methods and their iterative and automated implementation. As a case study, we discuss the application of the panel of these six methods to the transcriptome of *Plasmodium falciparum* intraerythrocytic developmental cycle. Assuming that spikes detected in this set have been labeled as artifacts by a biologist, in the second part of the chapter we discuss the effect of our smoothing procedure for different types of data analysis.

Keywords: malaria, DNA microarray, discrete mathematics, support vector machine (SVM), quality control

1. INTRODUCTION

To develop new drugs and vaccines that disable the malaria parasite *Plasmodium falciparum* (*P. falciparum*) [19], researchers need a better understanding of the regulatory mechanisms that drive the malarial life cycle. In [2], the first comprehensive transcriptome analysis of the *P. falciparum* asexual cycle, or intraerythrocytic developmental cycle (IDC), which is associated with the clinical symptoms of malaria, is provided. Data in [2] show that: (1) at least