A NEW ANNOTATION TOOL FOR MALARIA BASED ON INFERENCE OF PROBABILISTIC GENETIC NETWORKS

J. Barrera[1], R. M. Cesar Jr. [1], D. C. Martins Jr.[1], E. F. Merino[2] , R. Z. N. Vêncio[1] , F. G. Leonardi[1]
M. M. Yamamoto[2], C. A. B. Pereira[1], H. A. del Portillo[2]

[1]Institute of Mathematics and
Statistics, University of São Paulo
R. do Matão 1010
São Paulo, SP 05508-900, Brazil
Tel 55-11-3091- 6256
jb@ime.usp.br

[2]Institute of Biomedical Sciences,
University of São Paulo
Ave. Lineu Prestes 1374
São Paulo, SP 05508-900, Brazil
Tel 55-11-3091-7209
hernando@icb.usp.br

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 biochemically, had not been predicted. Recently, using global gene expression of the asexual blood stages of *P. falciparum* at 1h resolution scale and Discrete Fourier Transform (DFT) based techniques, it has been suggested that malaria parasites follow a rigid clock-wise program. Thus, a new list of coding genes with putative similar biological functions has significantly augmented new targets for vaccine and drug development. In this paper, genes are annotated under a different perspective: a list of functional properties is attributed to networks of genes representing subsystems of the regulatory expression system of *P. falciparum*. The model adopted 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 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 have been validated using the glycolytic pathway as a "gold-standard". Presently, we are trying to annotate genes not considered by the DFT approach.

Categories and Subject Descriptors

G.3. [**PROBABILITY AND STATISTICS**] Markov processes and statistical computing.

Subject Descriptor: probabilistic genetic networks

General Terms

Algorithms, Measurement, Documentation, Design, Reliability, Experimentation, Theory, Verification.

Keywords

Malaria, Annotation tool, Probabilistic genetic networks, Dynamical system, Markov chain, Mutual information.

1. INTRODUCTION

Malaria remains the most devastating parasitic disease worldwide, responsible each year for 300-500 million clinical cases and 1-2 million deaths, mostly in children below 5 years old [12]. Furthermore, the appearance of resistant parasite strains to most antimalarial drugs, the existence of insecticide-resistant *Anopheles* mosquitoes and the global environmental heating have exacerbated this 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 two years ago [8], but it has completely modified the way of thinking for the development of new vaccines, drugs and alternatives of control strategies. Moreover, it has allowed to initiate global scale studies on the transcriptome [2], proteome [7] and metabolome [14] of the parasite in different developmental stages.

Recent experimental evidence indicates that malaria parasites posses 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 [11], what suggests that these anti-sense transcripts should be involved in post-transcriptional regulation; reverse genetics approaches have shown that introns co-regulate expression of variant genes [4]; 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 [10].

Progressing the research effort, dynamical global gene expression measures of the asexual blood stages of the parasite at 1h-scale resolution were recently reported [1]. Moreover, using Discrete Fourier Transform (DFT) based techniques, the researchers verified that, during this life stage, the parasite seems to follow a rigid clock-wise program in which genes with common functions are transcribed at similar times. This study recognized 73% of the QC dataset (i.e., 3719 elements) expression signals with almost sinusoidal shape in the logarithmic scale or, equivalently, pulse like shape in a scale proportional to number of hybridized molecules. 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

Comment: % do que?
A porcentagem do total de genes da malaria.

comprehensive and biologically meaningful list of genes with putative similar functions [1]. Unfortunately, elements which had not almost sinusoidal shape and which represented 27% of the QC dataset (i.e., 1361 elements), were not included in these analysis.

In this paper a list of functional properties is attributed not to individual genes but to networks. To do so, it was created a tool that integrates mining of dynamical expression signals and conventional data basis (i.e., genoma, proteoma, metaboloma, and clinical data), with biological knowledge.

This annotation approach may be applied to all elements of the QC set, independent of the shape of their dynamical signals being sinusoidal or not. It consists in interpreting subsystems of the malaria expression regulation system as a probabilistic genetic network (i.e., a stochastic process that is a specialization of a Markov chain) [13], designing these networks from the dynamical signals observed and annotating the subsystems designed, using conventional data basis information and expert knowledge. The subsystems to be designed are defined from seed genes of particular biological interest, that is, the subsystems are composed by genes that predict or are predicted by seed genes [9]. For example, some genes analyzed by the DFT approach were used as seeds to discover other non sinusoidal genes associated with 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 adopted for designing a PGN. Section 4 describes software tools developed. Section 5 gives results of application of the design techniques to simulated PGNs and presents preliminary biological results obtained applying the proposed technique to the QC dataset [1]. Finally, in Concluding Remarks, the results and future steps of this research are discussed.

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 a lot of forward and feedback signals. These signals are RNA, produced by gene expression, and protein complexes, produced by interaction of proteins build by translation of mRNA. Protein complexes are feedback signals that control gene transcription and forward signals that, in the form of enzymes, control metabolic pathways. In such network, the expression of each gene depends on both its own expression and the expression values of other genes at previous instants of time. Due to this behavior, this complex network of interactions can 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 that model, we represent each gene by a variable which takes the expression value of that gene. All these variables, taken collectively, are the components of a vector called the *state of the system*. Each component (i.e., gene) of the state vector has associated a function that calculates its next value (i.e., expression value) from the state at previous instants of time. These functions are the components of a function vector, called *transition function*, that defines the transition from one state to the next and represents the gene regulation mechanisms. In order to formalize these ideas, we will introduce some definitions and notations. 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. This means that the transition function maps the present state to the next state. A finite dynamical system is given by, for every $t \geq 0$,

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

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

Systems defined as above are time *translation invariant*, that is, 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, where the stochastic transition function is a particular family of Markov chains, that is called probabilistic genetic network (PGN).

Consider a sequence of random vectors $X_0, X_1, X_2, \ldots$ assuming values in R^n and 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 $t \geq 1$,

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

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, for every *t* ≥ *1*,

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

where y is a realization of a random vector with distribution $p_{\bullet / x[t]}$.

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

i - $\pi_{Y/X}$ is homogeneous, that is, $p_{y/x}$ is independent of t.

ii - $p_{y/x} > 0$, for every states $x, y \in R^n$.

iii - $\pi_{Y/X}$ is conditionally independent, that is, for every states $x, y \in R^n$,

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

iv - $\pi_{Y/X}$ is almost deterministic, that is, for every state $x \in R^n$, there exists an state, $y \in R^n$ such that $p_{y/x} \approx 1$.

v – For every gene j there exits a vector a^j of integer numbers such that for every $x, z \in R^n$ and $y_j \in R$,

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

These axioms imply that each gene is characterized by a vector of coefficients a and a vector stochastic function g_j from Z, a set of integer 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 $x[t]$, that is, for every $t \geq 1$,

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

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

This model mimic the properties of a gene as a non linear stochastic gate and the systems built by compiling of these gates. In particular, the expression of a gene depends on a linear combination of excitatory and inhibitory input signals.

3. DESIGN OF PGNs

The goal of this research is to estimate a PGN [6] representing a subsystem of the malaria parasite gene expression network from dynamical microarray expression measures and biological knowledge. In the following, it is described the procedure adopted for PGN estimation.

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 diminishing its entropy. As a corollary of Property *ii*, the distribution with maximum entropy is the uniform distribution and the ones with minimum entropy are the ones with the whole probability mass concentrated in one point.

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

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

The mutual information is always positive or zero. 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

$$E[I(X,Y)] = H(Y) - E[H(Y/X)].$$

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 $x[t]$ pondered by an integer vector a, characteristic of gene y_i. For each vector a, with $a_i \in \{-1,0,+1\}$ and at most three values different of 0, the the mean mutual information is estimated. The first vectors a, that have greater mutual information are selected. These vectors indicate the connection between genes and the kind of connection, excitatory or inhibitory. At this moment, new vectors can be proposed by modifying selected vectors or just by adding new ones. The vectors modified or proposed will expand the list of selected vectors.

Using a selected vector for each gene, the architecture of a complete system is built. The transition matrix $\pi_{Y/X}$ is computed and the system is iterated till stability of π_t [3]. Then, we compute the entropy of π_t and choose the better ones according with smaller entropy. The motivation for this step is to choose simpler, more stable and robust systems. When the probability mass of π_t is concentrated, the system has a small set of states that are visited very often, what means that these states should be limited circles stable to perturbations. At this step, new systems may be proposed by choosing other systems not chosen by the low entropy criteria.

Finally, the chosen systems are ordered according to their likelihood, with respect to the observed data. The likelihood is computed for sequences of several sizes. A score is computed by the addition of the likelihood weighted by the size of the sequence considered in its computation. The better systems are the ones that have greater scores.

4. DEVELOPED SOFTWARE TOOLS

The designed software system estimates gene networks from dynamical expression measures and represent them as graphs linked to malaria data bases. A user-friendly graphical interface was implemented to facilitate the biological interpretation of the results. It uses GraphViz (http://www.research.att.com/sw/tools/graphviz/), a package to visualize graphs. This software receives files representing individual genes and their predictors as input, and generates a planar representation of the gene network. Moreover, for each node (i.e., gene) of the network, a color-code was assigned according to the functional biological categories defined in [1]: transcriptional machinery (pink), Cytoplasmic translation machinery (blue), glycolitic patways (yellow), etc. Each node has a link to a page with pointers to three public databases: PlasmoDB (http://plasmodb.org), Metabolic Pathways (http://biocyc.org/PFA/) and DeRisi's transcriptome database (http://malaria.uscf.edu/). Thus, this software allows easy access to different information of each target gene and will help in annotation of hypothetical proteins and null elements represented in the array. Another possibility is the generation of a set of individual subgraphs per gene, where each node points to a subgraph of its neighborhood. This facility permits navigability on the graph.

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 networks simulated have 12 genes that may be predicted from one to five genes or may even be independent. All network genes are ternary 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 1h-scale resolution observation of the asexual blood stages of *P. falciparum*). For each target gene, the five better pairs of predictors were computed according to the mutual information criteria. The *quality of a predictor g* was defined as the addition of the mutual information of all pairs 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 find at http://www.vision.ime.usp.br/camda04

5.2 Signal normalization and quantization

For validating the proposed methodology, the well known glycolytic pathway was studied. Before applying the predictor estimation techniques the signal was normalized and quantized.

Before quantization, the signals are normalized by the *normal transformation* η given by, for every random variable $g(t)$,

$$\eta[g(t)] = \frac{g(t) - E[g(t)]}{\sigma[g(t)]} ,$$

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}, respectively, down, null and up regulated in relation to the reference. The quantization of a gene signal g is performed by a *threshold mapping* given by, for every $t \geq 0$,

$$g(t) = \begin{bmatrix} +1 \; if \;\; s \geq h \\ 0 \;\;\; if \;\; l \leq s \leq h \\ -1 \; if \;\; s \leq l \end{bmatrix}$$

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

5.3 Glycolysis regulatory system prediction

The predictory capacity of the proposed model has been tested by choosing target genes that code for enzymes pertaining to the glycolytic pathway (hexokinase, phosphohexose isomerase PF14_0341, phosphofructokinase1 PF10755c, aldolase PF14_0425, triose phosphate isomerase PF14_0378, glyceraldehide 3 phosphate dehydrogenase PF14_0598, phosphoglycerate kinase PFI1105w, phosphoglycerated mutase PF11_0208, enolase PF10_0155). These genes have almost sinusoidal signal and were grouped in the ring state of the parasite life cycle according to the phaseogram [1].

In this prediction experiment all 5080 elements of the QC dataset were considered as possible predictors of the 8 glycolysis gene targets. For each target, it was computed the mutual information for the combination of all pairs of genes of the genome and the best five were chosen., that is, between 12,900,660 of gene pairs, the best five were chosen. Figure 1B shows three genes (id_oligo n132_136, j647_6 and c305) from the five best predictors of gene PFI0755c (id_oligo i13056_1). Note that in the best pair appears gene PF14_0341 (id_oligo n132_136), what agrees with the glycolysis metabolic pathway presented in Figure 1A. Besides note that the other gene of this pair, PF10_0097 (id_oligo j647_6), has a signal that is not sinusoidal as shown in Figure 1C.

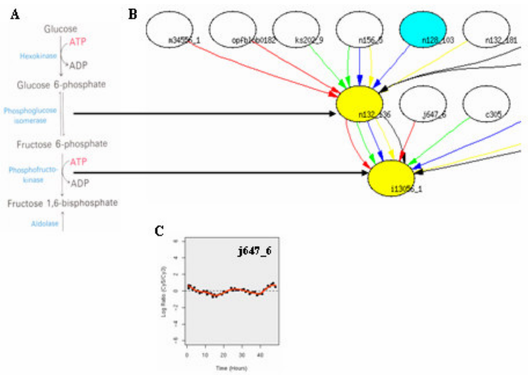

Figure 1. Predictory capacity of the PGN in Glycolisis. A. Initial steps of Glycolysis until the formation of Aldolase. B. Partial graphical interphase results displaying the best pair of combinations (red arrows) that predict phosphofructokinase. C. Dynamical expression of the gene PF10_0097 (id_oligo j647_6), non-sinusoidal and that was not included by the DFT approach [1].

The other targets were not predicted with the same precision but if the number of considered predictors increase they soon appear. The best 400 single predictors (i.e., just one gene predicting another gene) for each gene was calculated and this fact was verified for all 8 considered genes.

Note that the number of considered genes necessary to find the right predictor of a target gene is related to their positions in the phaseogram.

6. CONCLUDING REMARKS

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 QC dataset. Significantly, this technique was tested to target genes that code for enzymes of the glycolytic pathway and some of them (i.e. phosphofructokinase) were predicted with remarkable precision: the best pair of predictors, obtained through millions of different combinations, was formed by the expected gene (i.e., phosphoglucose isomerase) and a hypothetical protein, with non sinusoidal dynamical signals (Figure 1).

These preliminary 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 transform 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 deeper exploration of the malaria control system architecture. In particular, annotation of genes not considered by the DFT approach.

7. ACKNOWLEDGMENTS

Our thanks to Professor Bianca Zingales (IQ-USP) for first catalysing the interaction between JB and HAP, and to TDR-WHO for consodilating the interaction between the Departments of Parasitology and Computer Science through a Grant on Bioinformatics for Tropical Disease Research (http://malariadb.ime.usp.br/courses.html). The authors also thank FAPESP, CNPq and CAPES by their master, doctor, pos-doctor and research fellowships.

8. 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] Bozdech Z., Zhu J., Joachimiak M. P., Cohen F. E. , Pulliam B., and DeRisi J. L. Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. *Genome Biol. 4*, 2 (2003), R9.

[3] Brun, M.; Dougherty, E., and Shmulevich, I. Attractors in Probabilistic Boolean Networks: steady-state probabilities and classification. Submitted.

[4] Calderwood, M. S., Gannoun-Zaki, L., Wellems, T. E., andd 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), 34125-32.

[5] DeGroot, M. H. *Uncertainty, Information and Sequential experiments*. Annals of Mathematical Statistics, 3, 404-419, 1962.

[6] Dougherty, E. R., Bittner, M. L., Chen, Y., Kim, S, Sivakumar, K., Barrera, J., Meltzer, P., and Trent, J. M. *In Proceeding of Nonlinear filters in genomic control. IEEE-EURASI Workshop on Nonlinear Signal and Image Processing*. (Antalia, Turkey, 1999). 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] 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.

[9] 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.

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

[11] 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.

[12] Sachs, J., and Malaney, P. The economic and social burden of malaria. *Nature*, *415* (2002), 680-685.

[13] 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.

[14] 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.