Prediction of Novel Systems Components in Cell Cycle Regulation in Malaria Parasite by Subnetwork Alignments

Hong Cai,[1,*], Changjin Hong[2,*], Jianying Gu[3,*], Timothy G. Lilburn[4,*], Rui Kuang[2,§], Yufeng Wang [1,5§]
[1] Department of Biology, University of Texas at San Antonio, San Antonio, TX 78249, USA.
[2] Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA
[3] Department of Biology, College of Staten Island, City University of New York, Staten Island, NY 10314, USA
[4] Department of Bacteriology, American Type Culture Collection, Manassas, VA 20110, USA
[5]South Texas Center for Emerging Infectious Diseases, University of Texas at San Antonio, San Antonio, TX 78249, USA
Email addresses: HC: hong.cai@utsa.edu, CH: hong@cs.umn.edu, JG: jianying.gu@csi.cuny.edu, TGL: tlilburn@atcc.org, RK: kuang@cs.umn.edu, YW: yufeng.wang@utsa.edu
*These authors contributed equally to this work. §Corresponding authors

***Abstract*—With 300-500 clinical cases and 1-2 million deaths yearly, malaria contributes to enormous health care and economic burden worldwide. The advent of high throughput -omics technologies is driving new approaches to the identification of potential antimalarial targets. In this paper, we propose a neighborhood subnetwork alignment approach to uncover the network components involved in cell cycle regulation of the malaria parasite *Plasmodium falciparum* and to assign function to previously unannotated proteins.**

Keywords-malaria; systems biology;network; alignment; cell cycle

I. INTRODUCTION

One of the most ancient and devastating infectious diseases, malaria has continuous been a major contributor to morbidity and mortality in endemic regions. Every year 300-500 million people are infected with malaria, of which over one million die. The causative agents of malaria are protozoan parasites from the Genus *Plasmodium*, which infect primates, birds, or reptiles. Among the five species that infect humans, *P. falciparum* is the most lethal form.

Antimalarial drugs have been developed to kill the parasite or prevent its growth. However, in the past decades, the effectiveness of these drugs has been significantly reduced due to the evolution of parasites that are resistant to multiple drugs. Expectations that new targets could soon be identified have been raised by the completion of the genome sequences from *P. falciparum* [1], and by the subsequent high throughput studies [2-8]. These advances have created opportunities for a systems level interrogation of cellular networks associated with parasite development, survival, pathogenesis, and virulence.

Network alignment [9-16], one of the most popular methods in systems biology, however, does not seem feasible for the study of *Plasmodium* biology, simply because of the remote evolutionary relatedness of the parasite to other well studied model organisms. To tackle this problem, we recently developed a neighborhood subnetwork alignment algorithm, which is focused on the similarities between functional modules [17]. In this paper, we extend the subnetwork alignment approach to elucidate the systems components involved in cell cycle regulation, including several potential drug targets. An improved understanding of this important process will shed light on the fundamental mechanisms influencing parasite survival and pathogenesis.

II. METHODS

A. Subnetwork Querying

The problem of functional ortholog prediction was formulated as subnetwork querying, including the following steps: (1) the proteins related to cell cycle regulation (GO:0007049 : cell cycle) in *Escherichia coli* were mapped onto the PPI network. The neighbors of each cell cycle protein were selected to form the neighborhood subnetwork. (2) Each *P. falciparum* protein was mapped onto the parasite PPI network. (3) Neighborhood subnetworks were constructed to include the nearby neighbors. Because the size and density differ between the *E. coli* network and the *P. falciparum* network, a rule to control the neighborhood size was implemented: nearby proteins were selected by their distance to the central protein and the size of the neighborhood was restricted to 500 unless the central protein has more than 500 direct neighbors in the PPI network. Given a central protein p, let $N_k(p)$ denote the set of proteins that are k hops from p. The neighborhood of p is $N(p) = N_1(p) \cup N_2(p) \cdots N_k(p)$ such that $|N(p)| \leq 500$. Firstly, the neighboring proteins that were 1 hop from the central protein were included. If the size of the neighborhood was less than 500, we continued to include the proteins that were 2 hops from p. The hop distance was increased until the neighborhood size exceeded 500. (4) Each *E. coli* subnetwork was aligned against all the *P. falciparum* subnetworks. The central protein of the best-aligned *P. falciparum* subnetwork was identified as the functional ortholog of the *E. coli* cell cycle-related proteins.

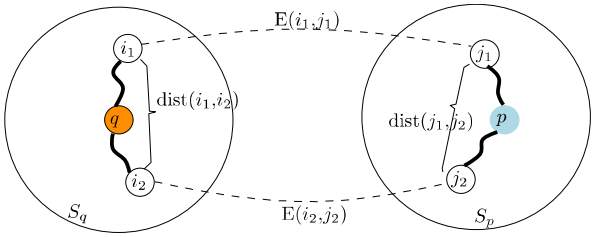

Figure 1. Subnetwork alignment. The alignment score between subnetwork S_p and S_q is the summation of the similarity scores between all pairs of matched shortest paths, (i_1, i_2) and (j_1, j_2), which is the based on the sequence similarity scores, E(i_1, i_2) and E(j_1, j_2), and the distances in the subnetworks, dist(i_1, i_2) and dist(j_1, j_2).

B. *Neighborhood Subnetwork Alignment with Graph Kernel*

Each possible alignment between *P. falciparum* and *E. coli* neighborhood subnetworks was scored and the similarity between these networks was summarized with a graph kernel [18, 19]. Given a pair of labeled graphs, a graph kernel is designed to summarize all possible isomorphic subgraphs (exact matches) in the two graphs. However, it is computationally infeasible to detect all isomorphic subgraphs due to an exponential number of subgraphs. A simplification is to compute the number of common paths between two graphs by a random walk on a product graph of the two compared graphs or by dynamic programming [20-22]. Alternatively, a graph kernel can also explicitly summarize the similarity between the shortest paths in the two graphs with each pair of shortest paths measured by a convolution kernel [23]. Since our focus is only on the paths that go through the central protein, we modified the shortest path graph kernel to only consider the paths between the central protein and the other proteins in the subnetwork. To capture the similarity of subnetworks, we focused on two important properties: the sequence similarity of the proteins and the role of the central proteins in the subnetwork. As shown in Figure 1, given two subnetworks S_p with central protein p and S_q with central protein q, a shortest path similarity function can be defined as

$$K(S_q, S_p) = \frac{1}{|S_q| + |S_p|} \prod_{\forall (i_1, i_2) \in S_q} B((i_1, i_2), S_p),$$

where,

$$B(i_1, i_2), S_p) = \max_{\forall (j_1, j_2) \in S_p} \frac{2E(i_1, j_1)E(i_2, j_2)}{dist(i_1, i_2) + dist(j_1, j_2)}.$$

E(x,y) is the E-value of the sequence alignment between proteins x and y, and dist(x,y) is the length of the shortest path connecting proteins x and y in the PPI subnetwork. In this similarity function, we took each pair of the proteins (i_1,i_2) in one subnetwork and identify the (j_1,j_2) in the other subnetwork that gives the maximum ratio between their sequence similarity with respect to (i_1,i_2) and the closeness in the subnetworks. Specifically, we computed the shortest path through the central protein between all pairs of proteins in the neighborhood subnetwork. The shortest paths of two neighborhood subnetworks are then compared and scored pairwise. The total of the alignment scores were reported as the subnetwork alignment score.

C. *Data and Network Analysis*

The *E. coli* PPI data were downloaded from the IntAct database [24]. The PPI data for *P. falciparum* were extracted from the STRING database [25]. Confidence scores (S) ranging from 0.15 to 0.999, were assigned for PPIs based on sequence similarity, pathway analysis, chromosome synteny, genome organization, phylogenetic reconstruction, and literature text mining. Cytoscape was used for network visualization [26]. NetworkAnalyzer was used to compute topological parameters of the networks, with the default settings. Gene Ontology (GO) enrichment analysis was conducted using BiNGO [27]. The hypergeometric test was used with the Benjamini and Hochberg false discovery rate (FDR) correction with a significance level of 0.05.

III. RESULTS AND DISCUSSION

A. *Neighborhood subnetwork alignments predicted 574 proteins that are associated with cell cycle regulation in malaria parasite.*

Malaria parasite has an atypical cell cycle that differs significantly from that in model eukaryotic organisms from yeast to mammals. There is no direct correspondence between schizogony during which parasite undergoes multiplication, and the typical G1, S, G2 and M phases of cell cycle in crown eukaryotes. Furthermore, the malaria parasite shows asynchronous nuclear divisions, organellar segregation, and morphogenesis of daughter merozoites during its cell cycle. Previously, to uncover the dynamics of the parasite cell cycle network, we developed a Variational Bayesian expectation maximization (VBEM) approach to infer regulatory relationships based on microarray time-series data [28]. In this paper, we attempted to identify the missing links in the cell cycle network using functional-module based subnetwork alignment.

We predicted that 574 proteins in *P. falciparum* were functional orthologs of known cell cycle proteins in *E. coli*. Over 34% of these predicted functional orthologs were annotated as "conserved *Plasmodium* proteins" or "putative uncharacterized proteins" of unknown function.

B. *Predicted protein set appears to be involved in important biological processes*

The predicted protein set is likely involved in diverse biological processes and molecular functions, as revealed by Gene Ontology Enrichment analysis (see Table I for representative functional categories). A putative cyclin (PFL1330c) may play an important role in cell cycle regulation. Yeast two-hybrid (Y2H) experiments [8] showed that it has physical interaction with an apical sushi protein (ASP) (Accession number PFD0295c), which has an adhesive “sushi” domain and is implicated in merozoite invasion process.

TABLE I. REPRESENTATIVE *P. FALCIPARUM* PROTEINS THAT WERE PREDICTED TO BE INVOLVED IN CELL CYCLE REGULATION.

Functional category	PlasmoDB Accession Number	Annotation
cyclin	PFL1330c	Cyclin-related protein, Pfcyc-2
cell differentiation	PFE0375w	cell differentiation protein, putative (CAF40)
chromosome organization	PFE0450w	Chromosome condensation protein, putative
	PF11_0062	Histone H2B
mitosis	PF13_0050	HORMA domain protein, putative
DNA repair	MAL7P1.145	Mismatch repair protein pms1 homologue, putative;
	PF10_0114	DNA repair protein RAD23, putative
	PF08_0126	DNA repair protein rad54, putative
DNA replication	PF07_0023	DNA replication licensing factor mcm7 homologue, putative
	PFL0580w	DNA replication licensing factor MCM5, putative
	MAL7P1.21	Origin recognition complex subunit 2, putative
Regulation of cell cycle	PF07_0047	AAA family ATPase, CDC48 subfamily (Cdc48)
	PFL1925w	Cell division protein FtsH, putative
protein phosphorylation	PFC0105w	Serine/threonine protein kinase, putative
	MAL13P1.278	Serine/threonine protein kinase, putative
	PF14_0294	Mitogen-activated protein kinase 1
	PF11_0464	Ser/Thr protein kinase, putative
	PF11_0239	Calcium-dependent protein kinase, putative
Proteolysis	PF14_0517	Peptidase, putative
	MAL13P1.184	Endopeptidase, putative
	PFL1635w	Ulp1 protease, putative
	PF10_0150	Methionine aminopeptidase
cytoskeleton	MAL8P1.146	filament assembling protein, putative
Heat shock	PFI0875w	Heat shock protein 70 (HSP70) homologue
	PFL0565w	Heat shock protein DNAJ homologue Pfj4
	PF11_0188	Heat shock protein 90, putative
	PFI0355c	ATP-dependent heat shock protein, putative
Pathogenesis	PFC0005w	Erythrocyte membrane protein 1, PfEMP1
	PFI0005w	Erythrocyte membrane protein 1, PfEMP1
	PFI1830c	Erythrocyte membrane protein 1, PfEMP1
Transcriptional regulation	PF10_0143	transcriptional coactivator ADA2 (ADA2)
	PFD0985w	AP2/ERF domain-containing protein PFD0985w
	PFL1085w	Transcription factor with AP2 domain, putative
	PF11_0442	Transcription factor with AP2 domain, putative
	PFE0840c	Transcription factor with AP2 domain, putative
	PF07_0126	Transcription factor with AP2 domain, putative
	PF10_0075	Transcription factor with AP2 domain, putative
	PFL1900w	Transcription factor with AP2 domain, putative

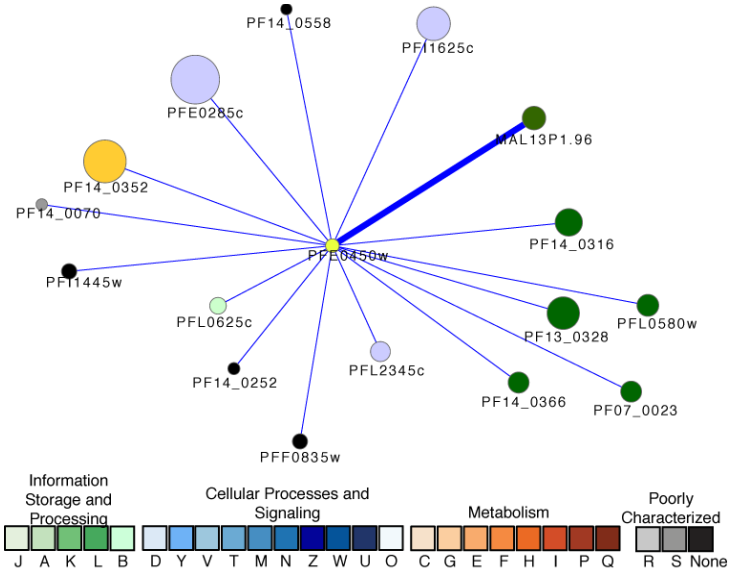

Figure 2. The proteins associated with a putative chromosome condensation protein PFE0450w. Node size is proportional to the degree of the connectivity of the node. Nodes are colored according to their functional classification in the eggNOG database.

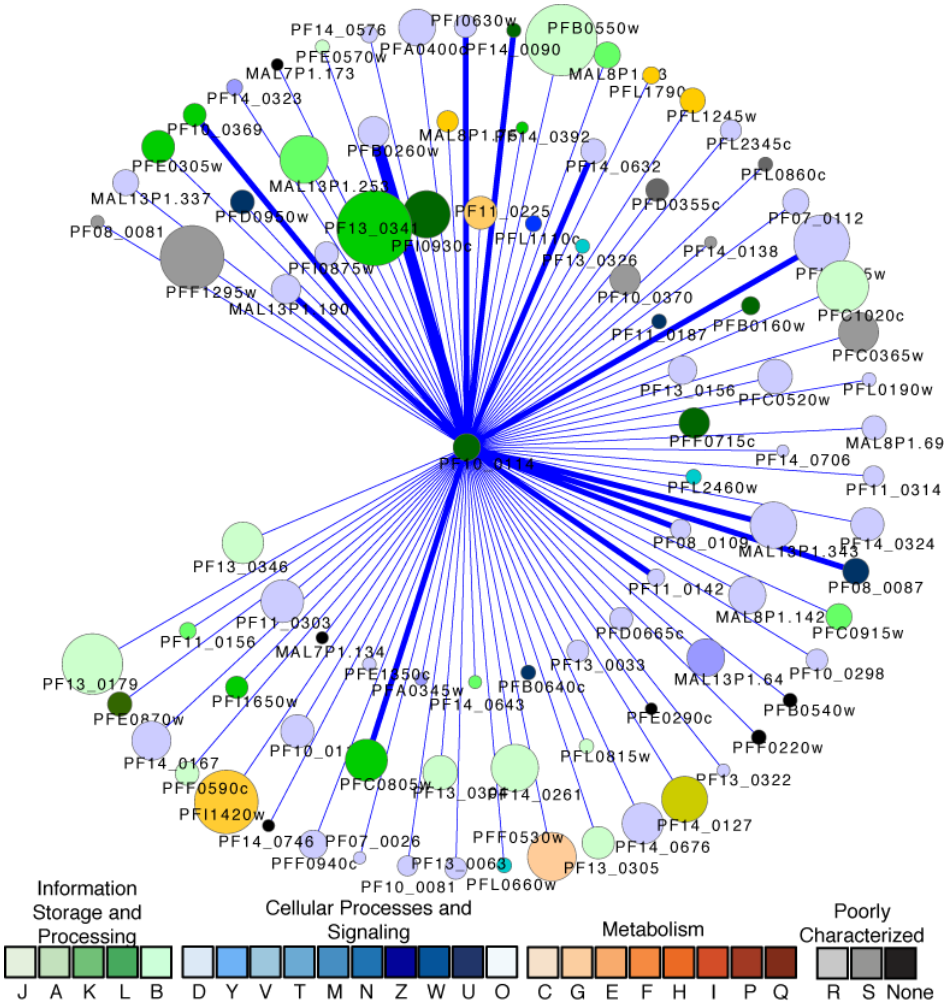

Figure 3. The proteins associated with a putative DNA repair protein RAD23 (PF10_0114)). Node size is proportional to the degree of the connectivity of the node. Nodes are colored according to their functional classification in the eggNOG database.

A number of other predicted proteins may be involved in cell division, mitosis, chromosome organization, and DNA replication. In addition to histone H2B, a putative chromosome condensation protein (PFE0450w), a member of the ATP-dependent chromatin remodeling complex [29], was predicted to be associated with cell cycle regulation. As shown in Figure 2, this protein has 16 protein-protein association partners, of which eight were verified by Y2H, including two tat-binding proteins pertinent to proteasome activites, a pre-mRNA splicing factor, an eukaryotic translation initiation factor 3 subunit 10, and three conserved *Plasmodium* proteins with unknown function. Most interestingly, it is associated with high molecular weight rhoptry protein 2 (RhopH2), which is localized in the rhoptries of schizonts and is implicated in cytoadherence and merozoite invasion to the red blood cell [30]. Several key components including DNA replication licensing factors and an origin recognition complex subunit were predicted by our subenwork alignment.

The cell cycle is also associated with genome integrity involving DNA repair mechanisms. A putative DNA repair protein RAD23 (PF10_0114) was predicted to have 92 protein-protein association partners, of which 22 were direct Y2H physical interactions. This protein is a member of an escort complex for proteasome-mediated degradation of non-native ER proteins. Its interactors include heat shock chaperone proteins, ATP-dependent proteases, serine-threonine kinases, and secreted proteins that have been implicated in stress responses, signaling cascades, and protein sorting and trafficking (Figure 3).

The complex nature of cell cycle is also manifested by its association with signal transduction. Various kinases and transcriptional regulators were predicted by the subnetwork alignments. PfMAP1 (PF14_0294) is a homolog of mitogen-activated protein kinase MAPK. This kinase may be related to parasite responses to a variety of exogenous or endogenous stimuli or environmental stresses. It is considered as a potential antimalarial target. At least seven parasite-specific ApiAP2 transcription factors were also predicted to be related to cell cycle regulation, underscoring the importance of transcriptional regulation. ApiAP2 proteins are becoming attractive drug targets due to their versatile roles in parasite life cycle and their distant evolutionary relationship to the host, indicative of little possible side-effects for humans [31].

CONCLUSIONS

A neighborhood subnetwork alignment approach was developed to predict the network components involved in cell cycle regulation. These network components are associated with a wide variety of biological processes ranging from DNA replication, DNA repair, transcriptional regulation, signal transduction, to stress responses, and pathogenesis. Many of these components have not been previously identified, thus demonstrating two significant abilities of our approach: it can identify the systems that drive important biological functions and it can assign potential functions to proteins for which homology-based approaches have failed.

ACKNOWLEDGMENT

We thank PlasmoDB for providing an all-in-one portal to the malaria genomic data. This work is supported by NIH grants GM081068 and AI080579 to YW. CH and KR are supported by University of Minnesota Grant-in-Aid of Research, Artistry and Scholarship. We thank the Computational Biology Initiative at UTSA for providing computational support. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences, National Institute of Allergy and Infectious Diseases, National Center for Research Resources, or the National Institutes of Health.

REFERENCES

[1] M. J. Gardner, N. Hall, E. Fung, *et al.*, "Genome sequence of the human malaria parasite Plasmodium falciparum," *Nature,* vol. 419, pp. 498-511, OCT 3 2002.

[2] Z. Bozdech, M. Llinas, B. L. Pulliam, *et al.*, "The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum," *PLoS Biol,* vol. 1, p. E5, Oct 2003.

[3] Z. Bozdech, S. Mok, G. Hu, *et al.*, "The transcriptome of Plasmodium vivax reveals divergence and diversity of transcriptional regulation in malaria parasites," *Proc Natl Acad Sci U S A,* vol. 105, pp. 16290-5, Oct 21 2008.

[4] K. G. Le Roch, Y. Zhou, P. L. Blair, *et al.*, "Discovery of gene function by expression profiling of the malaria parasite life cycle," *Science,* vol. 301, pp. 1503-8, Sep 12 2003.

[5] L. Florens, M. P. Washburn, J. D. Raine, *et al.*, "A proteomic view of the Plasmodium falciparum life cycle," *Nature,* vol. 419, pp. 520-526, OCT 3 2002.

[6] E. Lasonder, Y. Ishihama, J. S. Andersen, *et al.*, "Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry," *Nature,* vol. 419, pp. 537-542, OCT 3 2002.

[7] H. Ginsburg and L. Tilley, "Plasmodium falciparum metabolic pathways (MPMP) project upgraded with a database of subcellular locations of gene products," *Trends Parasitol,* Apr 12 2011.

[8] D. J. LaCount, M. Vignali, R. Chettier, *et al.*, "A protein interaction network of the malaria parasite Plasmodium falciparum," *Nature,* vol. 438, pp. 103-7, Nov 3 2005.

[9] S. Bandyopadhyay, R. Sharan, and T. Ideker, "Systematic identification of functional orthologs based on protein network comparison," *Genome Res,* vol. 16, pp. 428-35, Mar 2006.

[10] S. Bruckner, F. Huffner, R. M. Karp, *et al.*, "TORQUE: topology-free querying of protein interaction networks," *Nucleic Acids Res,* vol. 37, pp. W106-8, Jul 2009.

[11] J. Flannick, A. Novak, C. B. Do, *et al.*, "Automatic parameter learning for multiple local network alignment," *J Comput Biol,* vol. 16, pp. 1001-22, Aug 2009.

[12] M. Koyuturk, Y. Kim, S. Subramaniam, *et al.*, "Detecting conserved interaction patterns in biological networks," *J Comput Biol,* vol. 13, pp. 1299-322, Sep 2006.

[13] M. Koyuturk, Y. Kim, U. Topkara, *et al.*, "Pairwise alignment of protein interaction networks," *J Comput Biol,* vol. 13, pp. 182-99, Mar 2006.

[14] R. Kuang, E. Ie, K. Wang, *et al.*, "Profile-based string kernels for remote homology detection and motif extraction," *J Bioinform Comput Biol,* vol. 3, pp. 527-50, Jun 2005.

[15] R. Singh, J. Xu, and B. Berger, "Global alignment of multiple protein interaction networks," *Pac Symp Biocomput,* pp. 303-14, 2008.

[16] M. Zaslavskiy, F. Bach, and J. P. Vert, "Global alignment of protein-protein interaction networks by graph matching methods," *Bioinformatics,* vol. 25, pp. i259-67, Jun 15 2009.

[17] H. Cai, C. Hong, J. Gu, *et al.*, "Module-based Subnetwork Alignments Reveal Novel Transcriptional Regulators in Malaria Parasite Plasmodium falciparum," *BMC Systems Biology,* vol. Accepted, 2012.

[18] H. Kashima and A. Inokuchi, "Kernels for graphs," in *Kernel methods in computational biology*, B. Schölkopf, K. Tsuda, and J. P. Vert, Eds., ed: The MIT Press, 2004, pp. 155-170.

[19] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, *et al.*, "Graph Kernels," *Journal of Machine Learning Research,* vol. 11, pp. 1201-1242, Apr 2010.

[20] T. Gartner, P. Flach, and S. Wrobel, "On graph kernels: Hardness results and efficient alternatives," *Learning Theory and Kernel Machines,* vol. 2777, pp. 129-143, 2003.

[21] H. Kashima and A. Inokuchi, "Kernels for graph classification," in *The 2002 IEEE International Conference on Data Mining (ICDM 2002)*, 2002, pp. 31-36.

[22] H. Kashima, K. Tsuda, and A. Inokuchi, "Marginalized kernels between labeled graphs.," in *Proc. of the Twentieth International Conference on Machine Learning (ICML 2003)*, 2003, pp. 321-328.

[23] K. M. Borgwardt, H. P. Kriegel, S. V. Vishwanathan, *et al.*, "Graph kernels for disease outcome prediction from protein-protein interaction networks," *Pac Symp Biocomput,* pp. 4-15, 2007.

[24] S. Kerrien, B. Aranda, L. Breuza, *et al.*, "The IntAct molecular interaction database in 2012," *Nucleic Acids Res,* vol. 40, pp. D841-6, Jan 2012.

[25] D. Szklarczyk, A. Franceschini, M. Kuhn, *et al.*, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Res,* vol. 39, pp. D561-8, Jan 2010.

[26] M. E. Smoot, K. Ono, J. Ruscheinski, *et al.*, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics,* vol. 27, pp. 431-2, Feb 1 2010.

[27] S. Maere, K. Heymans, and M. Kuiper, "BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks," *Bioinformatics,* vol. 21, pp. 3448-9, Aug 15 2005.

[28] I. M. Tienda-Luna, Y. Yin, M. C. Carrion, *et al.*, "Inferring the skeleton cell cycle regulatory network of malaria parasite using comparative genomic and variational Bayesian approaches," *Genetica,* vol. 132, pp. 131-42, Feb 2008.

[29] C. R. Clapier and B. R. Cairns, "The biology of chromatin remodeling complexes," *Annu Rev Biochem,* vol. 78, pp. 273-304, 2009.

[30] L. Vincensini, G. Fall, L. Berry, *et al.*, "The RhopH complex is transferred to the host cell cytoplasm following red blood cell invasion by Plasmodium falciparum," *Mol Biochem Parasitol,* vol. 160, pp. 81-9, Aug 2008.

[31] H. J. Painter, T. L. Campbell, and M. Llinas, "The Apicomplexan AP2 family: integral factors regulating Plasmodium development," *Mol Biochem Parasitol,* vol. 176, pp. 1-7, Mar 2011.