20 Follador, I. *et al*. (2002) Epidemiologic and immunologic findings for the subclinical form of *Leishmania braziliensis* infection. *Clin. Infect. Dis.* 34, E54–E58
21 Seder, R.A. and Ahmed, R. (2003) Similarities and differences in $CD4^+$ and $CD8^+$ effector and memory T cell generation. *Nat. Immunol.* 4, 835–842
22 Costa, R.P. *et al*. (2003) Adhesion molecule expression patterns indicate activation and recruitment of $CD4^+$ T cells from the lymph node to the peripheral blood of early cutaneous leishmaniasis patients. *Immunol. Lett.* 90, 155–159
23 Bottrel, R.L. *et al*. (2001) Flow cytometric determination of cellular sources and frequencies of key cytokine-producing lymphocytes directed against recombinant LACK and soluble *Leishmania* antigen in human cutaneous leishmaniasis. *Infect. Immun.* 69, 3232–3239
24 Antonelli, L.R. *et al*. (2004) Antigen specific correlations of cellular immune responses in human leishmaniasis suggests mechanisms for immunoregulation. *Clin. Exp. Immunol.* 136, 341–348
25 Kaye, P.M. *et al*. (2004) The immunopathology of experimental visceral leishmaniasis. *Immunol. Rev.* 201, 239–253

1471-4922/$ - see front matter
doi:10.1016/j.pt.2005.06.007

Mining the malaria transcriptome

Manuel Llinás[1] and Hernando A. del Portillo[2]

[1]Department of Molecular Biology, Lewis-Sigler Institute for Integrative Genomics, 243 Carl Icahn Laboratory, Princeton University, Princeton, NJ 08544-1014, USA
[2]Departamento de Parasitologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, Avenida Lineu Prestes 1374, São Paulo, SP 05508-900, Brasil

Malaria remains the most devastating parasitic disease worldwide, and is responsible each year for >500 million infections and between one million and two million deaths of children under five years of age. *Plasmodium falciparum* is the most prevalent and deadly malaria parasite of humans, and a huge amount of data about it is now publicly available following completion of its genome sequence, the complete transcriptome of its asexual blood stages and proteomic analyses of its different life stages. Thus, new computational approaches are needed to analyze these data to yield biologically meaningful results that can be validated experimentally and, hopefully, lead to alternative control strategies. In this article, we highlight the importance of new computational approaches in mining the malaria transcriptome of the intraerythrocytic developmental cycle of *P. falciparum*.

An influx of large-scale datasets

Malaria continues to be one of the most devastating diseases to affect humans. With >500 million cases annually, including >2.5 million deaths, this disease causes a major economic and social burden in afflicted regions worldwide [Box 1(i)]. The full genome sequence for *Plasmodium falciparum* has recently been completed and is paving the way for large-scale genomic studies of this parasite [1]. Indeed, DNA microarray analyses that profile gene expression throughout the intraerythrocytic developmental cycle (IDC) of *P. falciparum* are already available [2–4] and transcript levels of several individual stage samples, including the gametocyte, sporozoite and merozoite, have been measured [5]. In addition, *in vivo* transcriptional profiling of the IDC from human patients has also been reported recently [6]. Furthermore, proteomic analyses of individual stages throughout the IDC are available, and global analysis of transcripts and protein levels are providing a preliminary link between gene expression and cellular protein translation [5,7,8]. Serial analysis of gene expression (SAGE) data are also available [9], more genome sequences from related *Plasmodium* species and isolates are completed or nearing completion [10,11] and further DNA microarray studies are imminent. The amount of data being generated is enormous, particularly considering that just one study of the 48-h IDC transcriptome comprises >350 000 data points.

With this large influx of post-genome data, malaria researchers are facing new challenges. How best to proceed? What tools are available to analyze these data? Do we proceed alone? Of course, these challenges are not

Box 1. Websites of interest

(i) World Health Organization malaria page (http://www.who.int/tdr/diseases/malaria/default.htm).
(ii) CAMDA meeting 2004 page (http://www.camda.duke.edu/camda04).
(iii) Queryable website from the DeRisi laboratory that hosts the IDC transcriptome [4] (http://malaria.ucsf.edu/).
(iv) MEME motif discovery algorithm for DNA and protein sequences (http://meme.sdsc.edu/meme/website/intro.html).
(v) AlignACE motif discovery algorithm (http://atlas.med.harvard.edu/).
(vi) PlasmoDB: the *Plasmodium* genome resource (http://plasmodb.org/).
(vii) KEGG – provides extensive metabolic pathway resources for many sequenced organisms (http://www.genome.jp/kegg/).
(viii) Malaria Parasite Metabolic Pathways – provides details of the established biochemical pathways that are specific to *Plasmodium* (http://sites.huji.ac.il/malaria/).
(ix) PlasmoCyc – queryable site that provides a graphical representation of metabolic pathways, protein-complex information and cellular-localization predictions (http://plasmocyc.stanford.edu/).
(x) Gene Ontology Consortium (http://www.geneontology.org/GO.doc.html).

Corresponding authors: Llinás, M. (manuel@genomics.princeton.edu), del Portillo, H.A. (hernando@icb.usp.br).
Available online 23 June 2005

unique to studies of *Plasmodium*; similar obstacles will be encountered during whole-genome studies of other parasitic protozoa of medical importance such as *Leishmania*, *Toxoplasma* and *Trypanosoma*. An obvious step is to accelerate the application of bioinformatics to tropical-disease research by using computational approaches to identify new biologically meaningful avenues of research. Ultimately, the goal is to develop alternative control strategies for malaria in the form of chemotherapeutics and/or vaccines. In this sense, several training workshops in bioinformatics for tropical-disease research are regularly offered worldwide. However, many malaria researchers are grappling with how best to use these datasets to benefit their own research.

Bridging the gap

At the fifth annual Critical Assessment of Microarray Data Analysis (CAMDA) meeting in 2004, another important step was taken towards addressing these issues [Box 1(ii)]. This meeting challenged delegates to apply existing methodologies, develop new algorithms and analyze computationally the comprehensive IDC transcriptome microarray dataset that is available to the *Plasmodium* community [Box 1(iii)]. The goal was to probe the data in an unbiased manner and attempt to identify *bona fide* starting points from which to direct further laboratory investigation*. The *P. falciparum* IDC transcriptome data were chosen because of their inherent complexity; unlike the cell cycle of yeast, in which there is a clear class of periodic cell-cycle genes in addition to co-regulated subgroups of genes, the majority of genes in the *Plasmodium* IDC is regulated in a periodic manner [3].

New directions from data mining

There are myriad questions posed by the available DNA microarray data that are well suited for rigorous computational examination. For example, owing to the high AT content (~90%) in the noncoding regions of the genome, predicting gene-regulatory regions in *Plasmodium* is a difficult task. However, combining established algorithms for detecting similarities in intergenic sequences – such as MEME [Multiple EM for Motif Elicitation, Box 1(iv)] and AlignACE [Aligns Nucleic Acid Conserved Elements, Box 1(v)] – with methods that focus on functionally related genes and/or genes with similar expression profiles will probably be more successful [5,7,12]. Furthermore, the absence of recognized regulatory transcription factors and a large number of potential RNA-binding proteins, the differential expression of ribosomal genes and the discrepancy of transcript and protein-abundance levels strongly suggest that control of gene expression in *Plasmodium* is unique and mostly posttranscriptional [13,14]. However, there do not seem to be many co-regulated sets of genes along the chromosomes in *Plasmodium* [3], which corroborates the idea that nuclear gene expression is not polycistronic [14]. Nonetheless, because the distribution of genes along chromosomes is often related to transcriptional regulation, chromosomal location coupled to gene-expression data can serve to identify regions of chromosomes that are actively transcribed, in addition to regions that are untranscribed. Thus, understanding the mechanisms that control global gene expression in *Plasmodium* could prove to be a key link to destabilizing this organism.

Throughout the intraerythrocytic development of *Plasmodium* species, many genes that are functionally unrelated are expressed in an extremely similar manner during the IDC and, therefore, cluster together by hierarchical clustering. In this case, the classic guilt-by-association method, which has been extremely successful at identifying co-regulated and functionally related genes in other organisms, falters. Therefore, new approaches are needed to take advantage of less obvious characteristics of the data. For example, investigators have further delineated relationships between genes using Bayesian decomposition [15], iterative signature algorithms [16], gene-expression networks [17] and probabilistic genetic networks [18]. These methods group related sets of genes by identifying similarities inherent within complex datasets. Other methods that focus on periodicity also reclassify the data, such as Lomb-Scargle periodograms [19] and the fast Fourier transform [3].

Two examples from the CAMDA meeting illustrate the value of using these more sophisticated mathematical and statistical approaches to explore DNA microarray data in malaria. In the first, subgroups of co-regulated *P. falciparum* genes were built using a model – probabilistic genetic network (PGN) – that was developed to represent genetic networks. PGN is a Markov chain with some additional properties that models the properties of a gene as a nonlinear stochastic gate. The subgroups (networks) are built by the coupling of these gates. A tool was then developed for integrating the PGN genetic networks with information from available *Plasmodium* databases and biological knowledge. The applicability of this tool for generating biologically meaningful PGN networks in malaria was demonstrated by successfully predicting a network for the genes of glycolysis. Subsequently, an apicoplast-specific PGN network was generated that revealed nuclear genes encoding putative apicoplast proteins that were not predicted by previous algorithms. Biological validation is now being sought.

In the second example from CAMDA, a statistical framework to examine spatial correlation between gene expression and location along chromosome regions was also designed. The goal was to examine spatially dependent co-regulation of gene expression during the 48-h IDC of *P. falciparum*. To do so, pairwise correlations between adjacent genes with or without distance restrictions, in addition to correlation through a formal covariogram function were performed using both the DNA microarray data and the *P. falciparum* annotated nucleotide sequence. The results, after accounting for intergenic distances, identified various spatially correlated regions along the *P. falciparum* chromosomes. These included small regions of highly correlated genes and regions of up to 100 kbp with only moderate correlations. This formal inferential method deals with spatial correlation, and sequence and expression data for the functional

* The results from CAMDA 2004 will be published in autumn 2005 by Springer as a collection of articles in a book – *Methods of Microarray Data Analysis V* – edited by Simon Lin and Patrick Hurban.

analysis of chromosomal regions with highly correlated genes on either the same or different DNA strands. Interestingly, the results revealed putative shared promoter and polycistronic regions in *P. falciparum*.

Associations and networks generated by these different methods can be strengthened further in multiple ways, and steps to integrate whole-genome datasets are underway. Using available resources such as PlasmoDB [Box 1(vi)], metabolic pathways from the Kyoto Encyclopedia of Genes and Genomes [KEGG, Box 1(vii)], Malaria Parasite Metabolic Pathways [Box 1(viii)] and PlasmoCyc [Box 1(ix)] can lend further confidence to predicted gene connectivities. In addition, parameters from the Gene Ontology Consortium [Box 1(x)] are useful for defining relationships between genes. Finally, well-established protein-interaction maps from other organisms such as yeast [20] suggest conserved interactions that might also exist in *Plasmodium*.

Future perspectives

This article is by no means an exhaustive account of all of the available bioinformatic methodologies for whole-genome microarray and proteomic-dataset analysis, and their application to *Plasmodium* research. Instead, our goal is to illustrate the fact that computational biology is fast becoming an essential tool for predicting and validating molecular mechanisms and biological networks of organisms such as *Plasmodium*. There are many large-scale datasets available for *P. falciparum*, including complete genome sequences, DNA microarray studies, proteomic analyses, expressed sequence tag (EST) data and SAGE data, and similar information is being generated for several other *Plasmodium* species. Incorporating all of the available data will help to derive important new relationships among genes. In the future, other meetings or forums that enable malariologists to interact with mathematicians, statisticians and bioinformaticians will be invaluable for providing further insight into the biology of *Plasmodium* species by applying more-sophisticated algorithms to the increasing quantity of available data.

References

1 Gardner, M.J. *et al*. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419, 498–511
2 Ben Mamoun, C. *et al*. (2001) Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite *Plasmodium falciparum* revealed by microarray analysis. *Mol. Microbiol.* 39, 26–36
3 Bozdech, Z. *et al*. (2003) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.* 1, E5
4 Le Roch, K.G. *et al*. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* 301, 1503–1508
5 Le Roch, K.G. *et al*. (2004) Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome Res.* 14, 2308–2318
6 Daily, J.P. *et al*. (2005) *In vivo* transcriptome of *Plasmodium falciparum* reveals overexpression of transcripts that encode surface proteins. *J. Inf. Dis.* 191, 1196–1203
7 Hall, N. *et al*. (2005) A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic and proteomic analyses. *Science* 307, 82–86
8 Florens, L. *et al*. (2002) A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* 419, 520–526
9 Patankar, S. *et al*. (2001) 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, 3114–3125
10 Carlton, J. *et al*. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* 419, 512–519
11 Carlton, J. (2003) The *Plasmodium vivax* genome sequencing project. *Trends Parasitol.* 19, 227–231
12 Militello, K.T. *et al*. (2004) Identification of regulatory elements in the *Plasmodium falciparum* genome. *Mol. Biochem. Parasitol.* 134, 75–88
13 Coulson, R.M. *et al*. (2004) Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*. *Genome Res.* 14, 1548–1554
14 Lanzer, M. *et al*. (1993) *Plasmodium*: control of gene expression in malaria. *Exp. Parasitol.* 77, 121–128
15 Moloshok, T.D. (2002) Application of Bayesian decomposition for analysing microarray data. *Bioinformatics* 18, 566–575
16 Ihmels, J. *et al*. (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics* 20, 1993–2003
17 Hasty, J. *et al*. (2001) Computational studies of gene regulatory networks: *in numero* molecular biology. *Nat. Rev. Genet.* 2, 268–279
18 Kauffman, S.A. (1969) Metabolic stability and epigenesist in randomly constructed genetic nets. *J. Theor. Biol.* 22, 437–467
19 Lomb, N.R. (1976) Least-squares frequency analysis of unequally spaced data. *Astrophys. Space Sci.* 39, 447–462
20 Yu, H. *et al*. (2004) Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. *Genome Res.* 14, 1107–1118

1471-4922/$ - see front matter
doi:10.1016/j.pt.2005.06.009

Genome Analysis

Genetic resistance to malaria in mouse models

Maria Hernandez-Valladares[1,2], Jan Naessens[1] and Fuad A. Iraqi[1]

[1]International Livestock Research Institute, Naivasha Road, PO Box 30709, Nairobi 00100, Kenya
[2]Institute of Molecular and Cell Biology of Africa, Naivasha Road, PO Box 30709, Nairobi 00100, Kenya

Murine models have proved to be excellent tools in the support of studies of the human genetic bases of malaria resistance and have enabled the mapping of 12 resistance loci, eight of them controlling parasitic levels and four controlling cerebral malaria. Further studies using this method have identified a Pklr variant that confers resistance to murine malaria, a result that shows the potential of this approach to aid the understanding of mechanisms of disease resistance. In the

Corresponding author: Hernandez-Valladares, M. (mhernandez@bxsoft.com).
Available online 20 June 2005