Title

The application of functional genomics, systems biology and drug development to the study of infectious diseases

Author

Zhu, Jingchun

Publication Date

2006

The Application of Functional Genomics, Systems Biology and Drug Development to the Study of Infectious Diseases

by

Jingchun Zhu

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

To Wyatt,
我的父亲，母亲，
and 姑姥爷.

Acknowledgements

I thank my thesis advisor, Dr. Joseph DeRisi for his guidance and support. His scientific vision, creativity and amazing energy have always been an inspiration for me. He often encouraged me to do science in a place that has not yet been explored and to not be afraid of learning new things that seemed foreign to me.

I would like to acknowledge several other people who made my thesis projects possible. The drug development project described in Chapter 5 was a close collaboration with Dr. Anang A. Shelat and Dr. R. Kiplin Guy. Using a large compound database they developed, Anang performed the virtual screens to search for anti-malarial compounds, predicted the "drug-like" properties for the derived compounds, and selected a subset of them for experimental validation.

During the course of the network modeling project (Chapter 4), I needed to perform several experiments to validate the computational predictions. I thank Dr. Ashwini Jambhekar for helping me to make knockout yeast strains, and in general at the bench. I thank Aaron Sarver for initiating the study of nitric oxide transcriptional response in *S. cerevisiae* and contributing his microarray data to the project.

I would like to thank Leslie Spector for her great help to proofread this thesis.

I also want to thank the other members of my thesis committee, Dr. Hao Li and Dr. Fred Cohen.

I am indebted to members of the DeRisi lab for their friendship and support throughout the years.

Finally, I want to acknowledge my family. The most important person I want to thank is my husband, Dr. Wyatt Tellis for his love and understanding, scientific discussions and programming expertise. I am also extremely grateful to have two great parents who let me to pursue my dreams. I would not have embarked on the career path if it were not for my gulaoye (grandpa), whose brilliance in math, physics and chemistry and dedication to work has inspired me since my childhood. At last, I thank Molly and Phillip for being the best in-laws a person can ask for.

The Application of Functional Genomics, Systems Biology and Drug Development to the Study of Infectious Diseases

Jingchun Zhu

Abstract

Genomics is creating a paradigm shift in the research of infectious diseases, transforming it from studying a few targets at a time to a genomic scale. We applied three genomic approaches to the study of malaria and its causative agents, a type of intracellular parasites belonging to the genus *Plasmodium*.

The first approach was to use DNA microarray technology to study the parasite transcriptome. The peculiarity of the *P. falciparum* genome made it difficult to produce a traditional cDNA probe-based microarray. We introduced a long oligonucleotide-based system in which each probe uniquely represents a single open reading frame and is optimized for other parameters including sequence complexity, secondary structure, and melting temperature. In order to produce such an optimal set of probes, we developed ArrayOligoSelector to automatically select gene specific long oligonucleotide probes for a complete genome.

In addition to study the parasite transcriptome, we developed a virtual drug development framework to identify anti-malarial compounds. The framework started with complete genome sequences and resulted in potential antimalarials; in the process it integrated a diverse and large amount of informatics data and computational methods. Using this framework, we identified 152 drug target genes by mining the phylogenomic patterns of 203 genomes and 77 co-ligands of those target proteins by comparative protein structure modeling and enzymatic predictions. Using the co-ligands as queries, we have computationally screened large compound collections to identify 1892 "drug-like" compounds that are structurally similar as well as commercially available.

Our third genomic strategy was to explore host pathogen interactions. We chose to focus on the pathogenic response to nitric oxide. Nitric oxide is an important mediator in the human innate immune response and a molecule associated with protection against severe malaria. The host innate immune response defends against infection by a wide range of pathogens including fungi and protozoan parasites. Since it is much easier to perform large-scale functional genomics experiments in a model organism other than in *Plasmodium*, we characterized the nitric oxide response in *S. cerevisiae* and applied a Bayesian network-driven approach to model the transcriptional response in that system.

Joseph

Dr. Joseph L. DeRisi
Dissertation Committee Chair

Table of Contents

List of Tables

List of Figures

Chapter 1 . Introduction

Malaria is one of the most deadly infectious diseases in the world, causing approximately 350- 500 million clinical episodes and as many as three million deaths annually [1-3]. Its causative agents are tiny intracellular protozoan parasites of the genus *Plasmodium*. This peculiar organism has a complicated life cycle that cycles between its mammalian host and the mosquito, its insect vector [4]. Although malaria as a disease has been documented in medical history for thousands of years [5], it was only until a little more than a hundred years ago that *Plasmodium* was discovered as the causative agent and the mosquito as the transmission vector [6]. Malaria has been virtually eradicated in the United States and most of the developed countries through mosquito control, yet the disease still plagues most of the developing world in the tropical and subtropical regions [2,3]. Although malaria has a tremendous impact on world health, it has been largely ignored by the pharmaceutical industry due to lack of financial interests and sometimes it has been referred to as a neglected disease.

Basic research using modern molecular approaches on *Plasmodium* has been limited due to difficulty in performing many of the standard genetic and biochemical techniques in the parasites (such as DNA recombination, transformation, and gene deletion), culturing the parasites (intraerythrocytic stages of *P. falciparum* can be grown in the laboratory, but the rest of the life cycle specimen must be salvaged from the mosquito vector or liver cells), divergence from any of the well-studied model organisms, and absence of its genome sequence.

In the past decade, advancement in genome sequencing and high-throughput measurement technologies has created a paradigm shift in biological research that transformed it from the study of a handful of targets at a time to a genomic scale. Fortunately for malaria research, the complete genomes of two *Plasmodium* species (*P. falciparum* and *P. yoelli yoelli*) became available in 2002 [7,8]. The genome sequences of its human host and mosquito vector *Anopheles gambiae* also became available between 2001 and 2004 [9,10]. With the newly available genome information, we can apply many of the new approaches developed in functional genomics, systems biology and computational drug development to the study of the malaria parasite [11].

Computational, mathematical and statistical methods play an increasingly important role in every aspect of this new biology. They are critical for designing the genome scale reagents that are necessary for performing the genomic experiments, analyzing large-scale data sets generated by this new biology, building models to interpret those data, and facilitating drug target identification and drug development.

Functional genomics is one approach to the study of gene function on a genome scale at different levels of cellular complexities including the transcriptome, proteome, and metabolome. To successfully perform functional genomics experiments, new kinds of reagents are required such as DNA microarrays, protein microarrays [12], gene deletion and small interference RNA (siRNA) libraries [13,14], genome-scale tagging libraries [15,16], and two-hybrid systems [17]. Generating these genome-scale reagents is not a simple task, in which robotic technologies, design algorithms, and a scalable experimental method are all critical components. In malaria research, most of the functional genomics effort has been focused on studying the parasite transcriptome, mostly because two critical technical difficulties were overcome: culturing the parasites in red blood cells, and the development of malaria genome microarrays [18-20].

The high (80%) AT content of the malaria genome posed a great obstacle to producing a traditional PCR probe-based microarray due to the high failure rate in PCR reactions (unpublished data, DeRisi lab). Long oligonucleotide-based DNA microarrays were proposed as an alternative, using a long probe sequence to represent each open reading frame [21]. An ideal set of probe sequences should uniquely represent a gene in the genome, avoid problematic sequence regions, and at the same time maintain a consistent melting temperature among all the probes. To design such a microarray, we developed a computational approach to automatically design gene-specific oligonucleotide genome array probes. The design algorithm and the accompanying software, ArrayOligoSelector, are described in Chapter 2.

Using ArrayOligoSelector (AOS), we designed a long oligonucleotide microarray for the complete genome of *Plasmodium falciparum*. This malaria chip has been used to study the transcriptome of the parasitic intraerythrocytic developmental cycle and sequence variation between different *P. falciparum* strains [18]. Chapter 3 presents the application of AOS on the *Plasmodium* genome to characterize the gene expression profile of the intraerythrocytic trophozoite and schizont stages of *P. falciparum*.

Ultimately, we want to take advantage of the genomic resource to develop new medicines to treat malaria. In the past, anti-malarial drug discovery was mainly focused on a small number of targets. Most the currently used drugs to treat and/or prevent malaria belong in four categories: quinine and its derivatives, antifolate combination drugs, artemisinin compounds, and tetracycline and its derivative antibiotics [22]. The completion of *Plasmodium*, human and other genomes has provided unprecedented opportunities for the discovery of new targets with novel modes of action.

For example, novel broad-spectrum drugs can potentially be identified by searching for genes that are conserved through evolution in bacteria, fungi and parasites but not present or exist in a very distinct form in the mammalian lineage [23].

In addition to genome sequence information, there are many resources and computational tools that can further lead us from drug target genes to identifying small molecule inhibitors. Compared to a decade ago, there is a much larger collection of functional, structural and chemical genomics databases [24-29], and more sophisticated computational methods for molecular docking of small molecules to protein targets [30,31]. With the development of combinatorial chemistry, large libraries of compounds have been synthesized and are available in a variety of chemical databases [26]. The pharmacokinetics properties (absorption, distribution, metabolism, and excretion) and toxicity of the compounds can be modeled using an *in silico* approach, which helps to decrease the late-stage failure in drug development [32-34]. The large number of available compounds also promoted the development of computational searching algorithms to retrieve compounds in chemical databases based on 2D or 3D structural similarities ([35]; reviewed in [36]).

Using these computational resources, a virtual anti-malarial drug development can be potentially carried out [37]. Chapter 5 describes a drug development framework we pursued to discover anti-malarial compounds *in silico*. This framework started with 203 complete genome sequences and resulted in 1893 potential antimalarials; in the process it integrated a diverse and large amount of informatics data of protein signatures and profiles, metabolic pathways, protein 3D structure models, and large compound collections. A large spectrum of computational methods was used or developed in this framework, which included sequence homology searches, ortholog identification, a phylogenomic analysis of a complete proteome, prediction for drug-like compounds, molecular fingerprints, a scoring function to systematically identify drug target proteins from the complete *P. falciparum* genome, and a virtual screen procedure composed of a two-step similarity search followed by "drug-like" properties and diversity filtering procedures. To evaluate this framework, a subset of the 1893 compounds will be tested by an *in vitro* malaria growth inhibition assay for enrichment of anti-malarial activities [38,39].

Although a virtual drug development strategy is associated with a large degree of uncertainty and false positives, an informatics method can still increase the possibility of promising compounds and therefore concentrate future drug development resources on such compounds [40-42]. More importantly, an *in silico* approach is much cheaper and faster. If it is coupled with high-throughput screening, much greater benefits can potentially be achieved. This can be particularly beneficial for battling many of the neglected infectious diseases such as malaria [37,43].

In addition to developing synthetic drugs to combat malaria, we can also explore the natural interactions between host and pathogen. For example, it has been hypothesized that the production of nitric oxide, a critical component of the human innate immune response, is an important mediator in the host's defense against *Plasmodium falciparum* malaria and is associated with protection against severe malaria [44-48]. However, the mechanism by which nitric oxide kills the parasite is still elusive [49,50]. A functional genomics study of the pathogenic defensive response to nitric oxide may shed light on the mechanism and potentially lead to the development of new anti-malarial therapies. Although genomic toolkits such as microarrays have become available for *Plasmodium*, it is still much easier to perform large-scale functional genomics experiments in model organisms such as *S. cerevisiae*. Host innate immune response defends against infection by a wide range of pathogens including fungi, bacteria, protozoan parasites and viruses [51-55]. Knowledge learned from studying the yeast response to nitric oxide can help us understand the defense mechanism used by fungi and other infectious agents against the human innate immune response [56].

Genomic surveys of nitric oxide triggered transcriptional responses have been carried out in several fungal organisms (*S. cerevisiae*, *Histoplasma capsulatum*, and *C. albicans*) [57-59]. These genomic experiments produced large datasets that were generated under various genotypes and experimental conditions. It is of great interest to mine these complete datasets and be able to decipher the relationships among environmental signals, genotypes, transcription factors and the corresponding transcriptional output. A network approach that can model multivariable systems is highly desirable. Chapter 4 describes a Bayesian network-driven approach to model the transcriptional response to nitric oxide in *S. cerevisiae*. We combined data mining, computational modeling, and experimental feedback in an iterative cycle of hypothesis generation and testing to build a network model to decode the relationship of nitric oxide, genotypes, and other environmental signals using the genome-wide transcriptional output as measured by microarrays. An automatic Bayesian network learning software, ExpressionNet, was developed and has been made freely available.

In summary, this thesis describes three major applications of genomics and computational methods to malaria research: designing genomic reagents, computational drug development, and a systems biology approach to decode transcription networks. They are a piece of the major genomic transformation in the research on infectious diseases. In the future, I expect to see a great expansion of mature molecular and genetic techniques in malaria research, continuing application of genomic technology and computational methods, and a greater degree of awareness of the disease impact on world health. And hopefully these will lead to the eradication of malaria.

Reference

1. Nabarro DN, Tayler EM (1998) The "roll back malaria" campaign. Science 280: 2067-2068.

2. Korenromp E (2005) MALARIA INCIDENCE ESTIMATES AT COUNTRY LEVEL FOR THE YEAR 2004 – PROPOSED ESTIMATES AND DRAFT REPORT -. In: Roll Back Malaria WHO, editor: World Health Organization.

3. WHO (2005) World malaria report 2005.

4. Sherman IW (1998) Malaria: parasite biology, pathogenesis, and protection. Washington, DC: ASM Press. xiii, 575 p. p.

5. Bruce-Chwatt LJ (1988) History of malaria from prehistory to eradication. In: Wernsdorfer W, McGregor I, editors. Principles and Practice of Malariology: Churchill Livingstone, Edinburgh, United Kingdom. pp. 1-59.

6. Garnham PC (1988) History of discoveries of malaria parasites and of their life cycles. Hist Philos Life Sci 10: 93-108.

7. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, et al. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. Nature 419: 512-519.

8. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498-511.

9. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science 298: 129-149.

10. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304-1351.

11. Sherman IW (2005) Molecular approaches to malaria. Washington, D.C.: ASM Press. xviii, 542 p., [518] p. of plates p.

12. Zhu H, Bilgin M, Snyder M (2003) Proteomics. Annu Rev Biochem 72: 783-812.

13. Giaever G, Chu AM, Ni L, Connelly C, Riles L, et al. (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418: 387-391.

14. Carpenter AE, Sabatini DM (2004) Systematic genome-wide screens of gene function. Nat Rev Genet 5: 11-22.

15. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, et al. (2003) Global analysis of protein localization in budding yeast. Nature 425: 686-691.

16. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, et al. (2003) Global analysis of protein expression in yeast. Nature 425: 737-741.

17. Tyers M, Mann M (2003) From genomics to proteomics. Nature 422: 193-197.

18. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, et al. (2003) The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. PLoS Biol 1: E5.

19. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, et al. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301: 1503-1508.

20. Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, et al. (2005) A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. Science 307: 82-86.

21. Rathod PK, Ganesan K, Hayward RE, Bozdech Z, DeRisi JL (2002) DNA microarrays for malaria. Trends Parasitol 18: 39-45.

22. Bloland PB (2001) Drug resistance in malaria. In: WHO/CDS/CSR/DRS/2001.4, editor.

23. Koonin EV, Galperin MY (2003) Sequence - evolution - function: computational approaches in comparative genomics. Boston: Kluwer Academic. xiii, 461 p., [411] p. of plates p.

24. Krummenacker M, Paley S, Mueller L, Yan T, Karp PD (2005) Querying and computing with BioCyc databases. Bioinformatics 21: 3454-3455.

25. Savchuk NP, Balakin KV, Tkachenko SE (2004) Exploring the chemogenomic knowledge space with annotated chemical libraries. Curr Opin Chem Biol 8: 412-417.

26. Dolle RE (2003) Comprehensive survey of combinatorial library synthesis: 2002. J Comb Chem 5: 693-753.

27. Yeh I, Hanekamp T, Tsoka S, Karp PD, Altman RB (2004) Computational analysis of Plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. Genome Res 14: 917-924.

28. Stuart AC, Ilyin VA, Sali A (2002) LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. Bioinformatics 18: 200-201.

29. Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, et al. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 32: D217-222.

30. Ewing TJ, Makino S, Skillman AG, Kuntz ID (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 15: 411-428.

31. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3: 935-949.

32. van de Waterbeemd H, Gifford E (2003) ADMET in silico modelling: towards prediction paradise? Nat Rev Drug Discov 2: 192-204.

33. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46: 3-26.

34. Ekins S, Nikolsky Y, Nikolskaya T (2005) Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. Trends Pharmacol Sci 26: 202-209.

35. Miller MA (2002) Chemical database techniques in drug discovery. Nat Rev Drug Discov 1: 220-227.

36. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, et al. (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. J Chem Inf Comput Sci 44: 1177-1185.

37. Maurer SM, Rai A, Sali A (2004) Finding cures for tropical diseases: is open source an answer? PLoS Med 1: e56.

38. Anderson MO, Sherrill J, Madrid PB, Liou AP, Weisman JL, et al. (2005) Parallel synthesis of 9-aminoacridines and their evaluation against chloroquine-resistant Plasmodium falciparum. Bioorg Med Chem.

39. Madrid PB, Wilson NT, DeRisi JL, Guy RK (2004) Parallel synthesis and antimalarial screening of a 4-aminoquinoline library. J Comb Chem 6: 437-442.

40. Polgar T, Baki A, Szendrei GI, Keseru GM (2005) Comparative virtual and experimental high-throughput screening for glycogen synthase kinase-3beta inhibitors. J Med Chem 48: 7946-7959.

41. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, et al. (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem 45: 2213-2221.

42. Paiva AM, Vanderwall DE, Blanchard JS, Kozarich JW, Williamson JM, et al. (2001) Inhibitors of dihydrodipicolinate reductase, a key enzyme of the diaminopimelate pathway of Mycobacterium tuberculosis. Biochim Biophys Acta 1545: 67-77.

43. Nwaka S, Ridley RG (2003) Virtual drug discovery and development for neglected diseases through public-private partnerships. Nat Rev Drug Discov 2: 919-928.

44. Gradoni L, Ascenzi P (2004) [Nitric oxide and anti-protozoan chemotherapy]. Parassitologia 46: 101-103.

45. Boutlis CS, Tjitra E, Maniboey H, Misukonis MA, Saunders JR, et al. (2003) Nitric oxide production and mononuclear cell nitric oxide synthase activity in malaria-tolerant Papuan adults. Infect Immun 71: 3682-3689.

46. Gyan B, Kurtzhals JA, Akanmori BD, Ofori M, Goka BQ, et al. (2002) Elevated levels of nitric oxide and low levels of haptoglobin are associated with severe malarial anaemia in African children. Acta Trop 83: 133-140.

47. Cramer JP, Mockenhaupt FP, Ehrhardt S, Burkhardt J, Otchwemah RN, et al. (2004) iNOS promoter variants and severe malaria in Ghanaian children. Trop Med Int Health 9: 1074-1080.

48. Cramer JP, Nussler AK, Ehrhardt S, Burkhardt J, Otchwemah RN, et al. (2005) Age-dependent effect of plasma nitric oxide on parasite density in Ghanaian children with severe malaria. Trop Med Int Health 10: 672-680.

49. Sobolewski P, Gramaglia I, Frangos J, Intaglietta M, van der Heyde HC (2005) Nitric oxide bioavailability in malaria. Trends Parasitol 21: 415-422.

50. Chiwakata CB, Hemmer CJ, Dietrich M (2000) High levels of inducible nitric oxide synthase mRNA are associated with increased monocyte counts in blood and have a beneficial role in Plasmodium falciparum malaria. Infect Immun 68: 394-399.

51. Fang FC (1999) Nitric oxide and infection. New York: Kluwer Academic/Plenum Publishers. xxv, 517 p. p.

52. Fang FC (2004) Antimicrobial reactive oxygen and nitrogen species: concepts and controversies. Nat Rev Microbiol 2: 820-832.

53. Missall TA, Lodge JK, McEwen JE (2004) Mechanisms of resistance to oxidative and nitrosative stress: implications for fungal survival in mammalian hosts. Eukaryot Cell 3: 835-846.

54. Nathan C, Shiloh MU (2000) Reactive oxygen and nitrogen intermediates in the relationship between mammalian hosts and microbial pathogens. Proc Natl Acad Sci U S A 97: 8841-8848.

55. Shiloh MU, Nathan CF (2000) Reactive nitrogen intermediates and the pathogenesis of Salmonella and mycobacteria. Curr Opin Microbiol 3: 35-42.

56. Vazquez-Torres A, Fang FC (1999) Therapeutic applications of nitric oxide infection. In: Fang FC, editor. Nitric oxide and infection. New York: Kluwer Academic/Plenum Publishers. pp. 475–488.

57. Nittler MP, Hocking-Murray D, Foo CK, Sil A (2005) Identification of Histoplasma capsulatum Transcripts Induced in Response to Reactive Nitrogen Species. Mol Biol Cell.

58. Hromatka BS, Noble SM, Johnson AD (2005) Transcriptional Response of C. albicans to Nitric Oxide and the Role of the YHB1 Gene in Nitrosative Stress and Virulence. Mol Biol Cell.

59. Sarver A, Derisi J (2005) Fzf1p Regulates an Inducible Response to Nitrosative Stress in Saccharomyces cerevisiae. Mol Biol Cell.

Chapter 2 . ArrayOligoSelector (AOS)

ABSTRACT

The complete genome sequence of an increasing number of organisms is becoming available. To exploit these new resources for the purpose of developing whole genome microarrays, we developed a program, ArrayOligoSelector (AOS), to systematically design gene-specific long oligonucleotide probes for entire genomes. For each open reading frame, the program optimizes the oligonucleotide selection based upon several parameters, including uniqueness in the genome, sequence complexity, probe secondary structure, GC content, and proximity to the 3' end of the gene.

Using AOS, we designed a long oligonucleotide microarray for the complete genome of *Plasmodium falciparum*, the most deadly causative agent of human malaria. This malaria chip has been used to study the transcriptome of the parasitic intraerythrocytic developmental cycle and sequence variation between different *P. falciparum* strains.

AOS is an open source program and is freely available for public use at http://arrayoligosel.sourceforge.net. AOS has also been used by scientists all over the world to design whole genome microarrays for many other organisms such as *S. cerevisiae*, *M. musculus* and *H. sapiens*.

The first section of this chapter presents the AOS design algorithm. The second section is the documentation for the program.

Part I. AOS Algorithms

Background

Two important technological advances have been instrumental in transforming biological research from the study of a handful of genes at a time to the age of genomics. The first is whole genome shotgun sequencing and assembly that allows complete genome sequences be obtained much cheaper and faster. As a result, the number of fully sequenced genomes, strains or individuals has increased dramatically. The second advancement is DNA microarray, a powerful technology that allows simultaneous measurements of gene expression for every gene in a whole genome, which has been used to gain important insights into processes such as development, responses to environmental perturbations, gene mutation, and host response to pathogens, and cancer [1-5].

To efficiently transfer genome sequence resources to functional genomics using microarrays, a new kind of reagent - whole genome microarrays – is needed. The traditional method for constructing a whole genome array was to generate PCR products for every gene in the genome, a laborious and time-consuming process with various rates of success. This became extremely challenging for genomes with very high AT content such as that of *P. falciparum* (80% AT). In addition, PCR probes have difficulty distinguishing genes with a high degree of sequence similarity. Oligonucleotide probe based platforms provide an alternative that overcomes these disadvantages [6,7]. The use of synthetic oligonucleotide probes eliminates the need for PCR. By carefully selecting probes from the unique regions, this platform provides a means to readily distinguish between genes that have a high degree of sequence similarity and avoid other problematic regions such as the various types of repetitive sequences or secondary structures.

Several competing platforms for producing oligonucleotide-based microarrays have emerged, differing in probe length, number of probes required per gene, nature of the production processes, design customization, and cost [7]. Affymetrix (Santa Clara, CA) pioneered the commercial market by producing high density GeneChips using photolithography and solid-phase DNA synthesis, on which each gene is represented by a set (~20) of short oligonucleotides (20 –25mer) [8]. Alternative to using chromium masks in conventional photolithography, NimbleGen's (Madison, WI) maskless arrays (24 – 70mers) are produced by light-directed synthesis of oligonucleotides controlled by a digital micromirror device [9]. Other commercial platforms include Agilent's (Palo Alto, CA) microarrays produced by an inkjet printing technology that synthesizes 60mer probes [10], CodeLink Bioarray$^{\text{TM}}$ (Amersham Biosciences, Piscataway, NJ) that uses a 3D polyacrylamide gel matrix as the slide surface for depositing 30mer oligonucleotide probes [11], and CombiMatrix's (Mukilteo, WA) CustomArray$^{\text{TM}}$ (50 - 70mers) which contains arrays of individually addressable microelectrodes for *in situ* oligonucleotide synthesis by means of an electrochemical reaction [12,13].

Commercial arrays are expensive, relatively difficult to customize probe design, and often limited to the model organisms. Spotted long oligonucleotide microarrays provide an inexpensive and highly customizable alternative. These arrays are produced in a similar fashion as the spotted cDNA arrays by depositing solutions of pre-synthesized oligonucleotide probes on a glass slide. The long oligonucleotide probes, usually 40 to 70mer in length, can be synthesized commercially. The array production can normally be performed in an in-house academic facility used for producing cDNA arrays, making it an ideal platform for academic laboratories.

Although spotted oligonucleotide arrays can be produced and used with a very similar method to those widely used for cDNA arrays, the success of oligonucleotide-based arrays are highly dependent on their probe design. To fulfill the objective of an oligonucleotide-based genome array, several design considerations need to be addressed. Most importantly, the probe sequence should be unique in the genome to minimize cross-hybridization. In addition, based on empirical rules used in primer designs, sequences that can form internal secondary structures should be avoided to maximize probe accessibility. Low complexity sequences should also be avoided to prevent nonspecific hybridization [14-16]. Other criteria are more unique to the design of a genome array, such as uniformity in probe melting temperatures and the proximity of probes to the 3' end of a gene. Another critical consideration is the choice of probe length, a balance between specificity and synthesis feasibility. In general, longer probes provide better specificity, but are associated with increasingly lower percentages of full-length probes (assuming 99% coupling efficiency, less than 50% of 70mer probes are full-length) and higher cost. Very short probes (<25mer) such as those used by GeneChip arrays require multiple probes per gene to improve signal specificity.

A computational approach is ideal to find the optimum design solution for this multi-parameter problem. Existing primer design programs are inadequate for designing a whole genome array. Therefore, we developed ArrayOligoSelector (AOS) specifically for the purpose of systematically selecting gene-specific long oligonucleotide probes for entire genomes. For each open reading frame (ORF), the program optimizes the oligonucleotide selection on the basis of several parameters, including uniqueness in the genome, sequence complexity, lack of self-binding, and GC content. Using AOS, we designed a long oligonucleotide microarray for the complete genome of *Plasmodium falciparum*, the most deadly causative agent of human malaria. This malaria chip has been used to study the transcriptome of the parasitic intraerythrocytic developmental cycle and sequence variation between different *P. falciparum* strains.

Similar approaches to oligonucleotide design have previously been described, but the exact algorithms, source code, and/or accompanying hybridization data are not available [8,10,17].

We made the algorithm, AOS source code and software, as well as the hybridization data publicly available to ensure public usage of the program, which is especially important for designing genome arrays for organisms like *Plasmodium* that hold minimal commercial interest, yet are immensely important for public health. Since we made AOS available, scientists all over the world have used AOS to design genome arrays for a wide variety of organisms including mouse, malaria, yeast and bacteria.

Algorithms

To design an optimum set of oligonucleotide probes for a given organism, AOS uses the ORF sequences and the complete genomic sequence as inputs, and then selects an optimum oligonucleotide for each ORF. The workflow of AOS consists of four major steps: 1) data preprocessing to ensure the correct sequence format and user inputs; 2) cognate sequence identification to discriminate true genomic targets from regions of potential cross-hybridization; 3) computing the following parameters for every oligonucleotide in an ORF sequence: uniqueness in the genome, internal secondary structures, GC percentage, and sequence complexity; 4) selecting a set of optimum oligonucleotide sequences using a rule-based filter procedure.

Step I: Data preprocessing

Correct data format, user inputs and computational resource are critical to ensure a smooth AOS execution. In the data preprocessing stage, AOS interacts with a user to obtain the sequence files (ORF sequences and the complete genome sequence), the oligonucleotide probe length, the choice for sequence masking, and the method to identify cognate sequences. It then verifies that the sequences are in the correct FASTA format, sequence identifiers do not contain white space characters to interfere with result parsing, no duplicated sequences in the sequence files, user input parameters are in the correct numerical range and selections, and the appropriate operating system is used. If all checks are passed, AOS is recompiled on the user's computer and proceeds to the next step.

Step II: Cognate genomic sequence identification

The cognate region is the genomic region where an ORF originates. Accurate identification of cognate regions is essential for differentiating true targets for an oligonucleotide from cross-hybridization regions. Since this information may not be easily available to all users, AOS opts to derive this information computationally based on the sequences provided in the two input files.

As the second step in the program workflow, AOS identifies an ORF's cognate region by reconstructing its exon structure. Each ORF was first aligned to the genomic sequences by a sequence homology search (BLAST or BLAT), alignments with 100% identical matches were stitched back together through a heuristic process to recapitulate the exon pattern. The principle behind this strategy is that individual exons should be among those perfect alignments, and the goal is to identify those specific perfect alignments and the exact order they should be arranged in to form the corresponding ORF. However, the difficulty comes from the fact that not every 100% identical alignment region is necessarily a part of the ORF exon structure. Although an exhaustive search of all possible arrangements of any number of perfect alignment regions can find the correct exon structure, the number of arrangement combinations increases factorially as the number of perfect alignments increases($\sum_{n} n!$), which makes an exhaustive strategy impossible to complete if a great number of perfect alignments was initially identified.

Therefore a heuristic approach is used to decrease the search space. The first heuristic trick is that a search can only start from either a perfect alignment of >50 bp, or a "must-use" alignment (see below for details). Secondly, a search can only continue by adding other perfect alignment regions that satisfy the following spatial constraints: same chromosome and strand orientation; proximity to all existing alignment regions (<3000 kb); minimum overlap with existing alignment segments (<70% of the smaller of the new and existing regions); consistent arrangement in both ORF and genome sequences (e.g. if the new region was to the 5' end of an existing region in the ORF sequene, it must be so in the genomic sequence as well). Third, a seach stops when the sum of the existing regions reaches the length of the ORF. Fourth, a search also stops when the sum of all potential regions is highly unlikely to reach the length of the ORF. Fifth, only 100% perfect alignment regions can be considered (SNPs not allowed). Sixth, if the first high scoring hit alignment (best alignment) to any chromosome is less than 50bp, alignment regions from the entire chromosome will not be considered.

In simple terms, the AOS search procedure is to construct combinations of alignment regions; each combination a solution for the correct exon pattern. Procedurely, the above heuristics is implemented as first identifying all perfect alignments, followed by building a connectivity matrix to specify compatible alignments if an alignment has already been selected as part of an exon pattern (based on the spatial constraints described above: compatible chromosome, strand direction, proximity, overlap, and spatial orientation). Subsequently, AOS identifies the "must-use" alignments by scanning for regions in the ORF sequence that are covered by a single perfect alignment, and the corresponding alignment is referred to as the "must-use" alignment. After that, the AOS search starts to construct a list with a single alignment that is either a "must-use" alignment or a perfect alignment >50 bp. AOS proceeds to add additional perfect alignments (n) that are allowed by the connectivity matrix. The original list is duplicated n times and a different alignment is added at the end of each list. This duplication and extension procedure continues until when existing alignments in a list have reached the full length of the ORF. If existing alignments in a list plus all their potential additions (allowed by the connectivity matrix) cannot reach the full length of the ORF, the list is eliminated from furthur consideration. At the end of the search process AOS finds a collection of lists; each contains one or more alignments. Each list is a possible solution for the real exon pattern.

To ensure the accuracy of the results, lists in the final collection are re-examined. Only combinations within ±20 bp of the ORF full length size and able to generate the original ORF sequence in a correct order are kept as solutions for the exon pattern reconstruction. Multiple solutions are allowed. The corresponding exon locations in the genomic sequences are extracted as an ORF's cognate region. This cognate region information is stored in disk to be used in the uniqueness calculation in Step III.

Users can choose to use either the BLAST or BLAT program for sequence alignment to identify the perfect alignment regions [18]. BLAST is more sensitive and typically generates a greater number of alignments, therefore resulting in a bigger search space and slower speed for exon pattern reconstruction. Using BLAT is faster, but it has the risk of missing short alignments. It is important to note that the low complexity filter must be turned off during alignment at this step, otherwise cognate regions will fail to be identified. However, this is at a great cost of computational speed due to the large number of short low complexity alignments generated.

Step III: Parameter computation

In the parameter computation step, AOS calculates values for the following features for every oligonucleotide sequence: uniqueness in the genome, internal secondary structure, sequence complexity, GC percentage, and position in the ORF. Each feature is computed using an independent module, which can also be used as a stand-alone program to obtain individual parameter. The parameter values were written to disk for use in the later selection step.

1. Uniqueness in genome

The uniqueness of an oligo in the genome was measured as the theoretical binding energy of the worst potential cross-hybridization to its homologous regions in the genome. Potential cross-hybridizations are detected by BLASTN alignment, followed by binding energy calculation using the energy module. The uniqueness score of an oligonucleotide is the most stable binding energy between the oligo and the genome excluding the corresponding ORF's cognate region.

In earlier versions of AOS, we used the number of sequence identity in BLAST alignment between the oligo sequence and the genomic cross-hybridization targets as our measurement of cross-hybridization. But our experimental results demonstrated that this metric was a poor predictor for cross-hybridization of different hybridization binding structures. A DNA-DNA duplex becomes less stable when bulges (sequence mismatches) are introduced into the middle of the duplex. Given the same number of perfect base pairing (sequence identity), hybridization signal strength is stronger when the matches form a continuous stretch compared to a different duplex structure with mismatches distributed in the middle (Figure 3-4).

To overcome the difficulty to predict cross-hybridization by simple sequence identities, we implemented the energy module to calculate hybridization binding energy, in order to unify predictions of different binding structures into a single formulation. The binding energy calculation is based on the nearest neighbor model for calculating nucleic acid helix formation and melting temperatures [19], RNA secondary structure prediction algorithms [20,21], and experimentally estimated thermodynamic free energy parameters for oligonucleotide duplexes and RNA secondary structures [22-28].

In addition to careful modeling of the duplex energetic property, the accuracy of the binding energy calculation is highly dependent on initial accurate identification of those DNA duplexes. AOS uses BLASTN alignment program to identify those regions between the ORF and the genome, and then uses the energy module to calculate the binding energy between the aligned regions.

1.1 Binding energy score

The binding energy score is the summation of the following three terms: the base pair stacking energy between the two adjacent base pairs (such as dAA/dTT), the initial binding energy required for helix initiation, the interior and bulge loop destabilizing energy.

The base pair stacking energy is derived based on the nearest neighbor rules, i.e., the energy of the duplex is the addition of free energy terms of each adjacent Watson-Crick base pair, which includes energy contribution for both base pair stacking and hydrogen bonding. For example, in the following five base pair duplexes, the first two base pairs (dAT/dAT) have a stacking energy of –0.9 kcal/mol, the second and third base pairs (dTT/dAA), –1.2 kcal/mol; the third and fourth base pair (dTG/dCA), –1.5 kcal/mol and so on. The final stacking energy term is (–0.9) + (–1.5) + (–1.2) + (-2.3) = -5.9 kcal/mol.

```
A T T G C
| | | | |
T A A C G
```

Individual stacking energy parameters were obtained by experimentally estimating nearest neighbor parameters for all ten adjacent base pair combinations [22].

The helix initiation energy term models the free-energy change for initiation of DNA duplex, which was estimated experimentally to be +3.4 kcal/mol [22,28].

The interior loop or bulge loop can form when mismatches are closed by at least 2 base pairs. Mismatches on both strands result in the formation of an interior loop. If a mismatch only exists on one strand, the formation is an interior bulge. Both interior loop and bulge contributed destabilizing free energy to the duplex. The loop or bulge destabilizing energy is modeled as the sum of the following three terms: an entropic term that depends on the size of the loop or bulge; terminal stacking energy for the mismatch base pairs adjacent to both closing base pairs, which sometimes provides a favorable free energy; an asymmetric loop penalty for non-symmetric interior loops [20]. The terminal mismatch stacking energy parameters such as dAA/dTA (+0.61 kcal/mol) were estimated experimentally using short nucleic acid duplexes [23-27]. The parameters for the entropic term were derived from parameters used in RNA secondary structure prediction, which were empirical approximations of experimental measurements (Table 2-1) [21].

The parameter for asymmetric loop penalty was based on a study of internal loops in oligonucleotides by Peritz et al. [29,30]. An asymmetric internal loop with a size of $N1$ and $N2$ nucleotides should be penalized by $N*f(M)$ kcal/mol, where $N= | N1 - N2 |$, M is the minimum of 5, $N1$ or $N2$, and $f(1) = 0.7, f(2) = 0.6, f(3) = 0.4, f(4) = 0.2$ and $f(5) =$ 0.1.

The nearest neighbor model had good agreement with experimental data on short duplexes. It is well known that the binding free energy and melting temperature of double-stranded DNA molecules plateau at a longer length. However, evidence for size limitation of the nearest neighbor model and parameters is sparse. In addition, the above thermodynamic parameters used in our binding energy calculation were estimated from experimental measurements on short oligonucleotide duplexes (<20 bp). Therefore, although we used both to model long oligonucleotide duplex binding stability, the binding energy values should be viewed as a function of binding stability on a relative scale, rather than be interpreted as the absolute free energy generated during DNA duplex formation.

1.2 Energy score correlates linearly with measured hybridization strength on 70mer oligonucleotides

Although the energy module and parameters are probably not an accurate depiction of the true binding energetic property of long oligonucleotide DNA duplexes, we were interested in using the energy score as a relative measurement of hybridization strength, which could then be used to estimate potential cross-hybridization. To evaluate the utility of the binding energy score to measure cross-hybridization, we conducted experiments on a series of 70mer oligonucleotides with various predicted duplex structures.

We designed several series of 70mer microarray probes that target the *Plasmodium falciparum* genome. In each series, there was a perfect 70mer that matched the coding sequence of an ORF perfectly; the rest of the series was composed of 70mers with various numbers of mismatched base pairs distributed either at the terminals or in the middle of the 70mer. We hybridized transcripts extracted from various stages of *P. falciparum* parasites and then obtained the relative hybridization signal of the mismatched 70mers to the perfect 70mer in each series. The binding energy score of each mismatched 70mer was computed for the duplex (alignment between the mismatched and perfect 70mers). Results demonstrated that there existed a linear relationship (Pearson correlation coefficient $r = -0.91$) between binding energy scores and the relative hybridization strength (Figure 3-3).

1.3 Speed Optimization

Binding energy scores are calculated as the sum of many independent terms, such as the base pair stacking energy and loop destabilizing penalties. Therefore, for two adjacent oligonucleotide probes (with a single base pair offset), their energy score calculation involves a large degree of redundancy. In addition, potential cross-hybridization regions were initially identified by BLAST, followed by the binding score calculation, if we simply used a single oligonucleotide sequence as the input to the energy module, essentially the same BLAST operation would be carried out for adjacent oligonucleotides as well. Both kinds of redundancy would dramatically decrease the speed of the energy module.

To increase the speed, we optimized the energy module by the following strategies. First, we only performed a single BLAST alignment using the entire ORF sequence. Second, we computed the free binding energy score for an entire alignment instead of for a single oligonucleotide, excluding any alignment from the cognate sequence region. Third, in addition to a single energy score, we recorded the score contributions from every adjacent base-pair in the entire alignment. To derive the binding energy score for an oligonucleotide, we simply summed up the score contributions from the corresponding regions in the alignment. Since an oligonucleotide sequence could be covered by more than one alignment, the final binding energy score was the most stable energy score (the largest absolute value).

2. Internal secondary structure

The secondary structure module measures the potential of forming an internal hairpin structure within an oligonucleotide. A fast approximation for detecting internal hairpins is by aligning the oligo sequence with its reverse compliment. We implemented the Smith-Waterman algorithm to search for the optimal local alignment [31] and used the alignment score to represent the potential to form internal hairpins. PAM47 DNA matrix is used (match +5, mismatch –4, gap opening –7, gap extension 0) in the implementation for local alignment.

Sophisticated RNA secondary structure prediction methods such as Mfold were available [32] and likely to generate more accurate results, but they are much slower computationally.

3. Sequence complexity

The sequence complexity module measured the level of oligo sequence complexity using the LZW compression algorithm [33]. The advantages of this method are fast computational speed and no need for prior information for low complexity sequence elements. It is implemented as the size of the oligonucleotide sequence minus its compressed version in bytes.

4. GC content

GC content is a key factor determining DNA duplex melting temperature. We used it as the proxy for melting temperature, calculated as the the number of G C base-pairs over the length of the oligo.

Step IV: Optimum selection

The last step of the AOS algorithm is to select a set of optimum oligonucleotide sequences based on the parameters computed in Step III. An ideal oligo probe has a small negative value of binding energy score (unique in the genome), a small secondary structure score (lack of internal hairpins), a small sequence complexity score, a %GC close to the user-defined target %GC, and close to the 3'end of the ORF sequence.

We implemented a rule-based filtering procedure to select for the optimum oligonucleotide. The first filter is the uniqueness filter. Oligos belonging to a single ORF are ranked first by their uniqueness scores (binding energy score). Oligos scoring better than both an optional user-defined threshold and the default cutoff are kept in the candidate pool. The default cutoff is defined as the larger (smaller absolute value, note energy scores were negative values) of the following two terms: the 5th percentile in the rank, and the best uniqueness score minus 5 kcal/mol.

The second filter is to eliminate any oligos with user-defined (optional) unwanted sequences, such as a long stretch of AT sequence.

The third filter operates on the sequence complexity parameter and secondary structure score in parallel. Similar to the operation on energy scores, oligos that pass the cutoffs can proceed further. Although it only operates on the current candidate pool (oligos that passed the previous two filters), the cutoffs are determined using the complete set of oligos belonging to a single ORF. The initially cutoffs are determined at the top 33rd percentile of the rank by either the secondary structure score or the sequence complexity score. If there is no candidate oligo that can pass both thresholds simultaneously, each cutoff is relaxed incrementally (secondary structure score cutoff increases by 10; sequence complexity score cutoff increases by 1) until one or more oligos pass both thresholds simultaneously.

The fourth filter operates on the %GC parameter. Initially, only oligos with the user-defined target %GC can pass. If no oligo in the current candidate pool satisfies this criterion, the %GC boundaries are relaxed by 1 percentage point at a time in each direction until one or more oligos score within the range.

The final filter operates on the 3' proximity to select the oligo that is closest to the 3' end of the parent ORF. This oligonucleotide is our optimum selection. At this point, AOS reaches its final step to generate program output of the optimum oligo selection.

Occasionally, if a user wants to design more than one oligo per ORF, AOS will attempt to select non-overlapping (must be >10 bp apart at the oligo starting positions, but typically >50bp) oligos from the current pool. If this is not successful using the current candidates, the selection procedure iterates from the combined secondary structure and sequence complexity filter to the 3' proximity filter, until the desired number of oligos is selected, or when the cutoffs are fully relaxed and the candidate pool reaches its maximum size.

Reference

1. DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-686.

2. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. (2000) Functional discovery via a compendium of expression profiles. Cell 102: 109-126.

3. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503-511.

4. Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, et al. (2005) The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. Nat Genet 37: 544-548.

5. Rubins KH, Hensley LE, Jahrling PB, Whitney AR, Geisbert TW, et al. (2004) The host response to smallpox: analysis of the gene expression program in peripheral blood cells in a nonhuman primate model. Proc Natl Acad Sci U S A 101: 15190-15195.

6. Hardiman G (2003) Microarrays Methods and Applications; Hardiman G, editor. Eagleville: Dna Press.

7. Hardiman G (2004) Microarray platforms--comparisons and contrasts. Pharmacogenomics 5: 487-502.

8. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14: 1675-1680.

9. Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, et al. (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res 12: 1749-1755.

10. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, et al. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol 19: 342-347.

11. Ramakrishnan R, Dorris D, Lublinsky A, Nguyen A, Domanus M, et al. (2002) An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. Nucleic Acids Res 30: e30.

12. Tian J, Maurer K, Tesfu E, Moeller KD (2005) Building addressable libraries: the use of electrochemistry for spatially isolating a heck reaction on a chip. J Am Chem Soc 127: 1392-1393.

13. Tesfu E, Maurer K, Ragsdale SR, Moeller KD (2004) Building addressable libraries: the use of electrochemistry for generating reactive Pd(II) reagents at preselected sites on a chip. J Am Chem Soc 126: 6212-6213.

14. Chavali S, Mahajan A, Tabassum R, Maiti S, Bharadwaj D (2005) Oligonucleotide properties determination and primer designing: a critical examination of predictions. Bioinformatics 21: 3918-3925.

15. van Baren MJ, Heutink P (2004) The PCR suite. Bioinformatics 20: 591-593.

16. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.

17. Rouillard JM, Herbert CJ, Zuker M (2002) OligoArray: genome-scale oligonucleotide design for microarrays. Bioinformatics 18: 486-487.

18. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664.

19. Tinoco I, Jr., Uhlenbeck OC, Levine MD (1971) Estimation of secondary structure in ribonucleic acids. Nature 230: 362-367.

20. Lyngso RB, Zuker M, Pedersen CN (1999) Fast evaluation of internal loops in RNA secondary structure prediction. Bioinformatics 15: 440-445.

21. Turner DH, Zuker M (2005) Free Energy and Enthalpy Tables for RNA Folding.

22. Sugimoto N, Nakano S, Yoneyama M, Honda K (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. Nucleic Acids Res 24: 4501-4505.

23. Peyret N, Seneviratne PA, Allawi HT, SantaLucia J, Jr. (1999) Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches. Biochemistry 38: 3468-3477.

24. Allawi HT, SantaLucia J, Jr. (1998) Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: sequence dependence and pH effects. Biochemistry 37: 9435-9444.

25. Allawi HT, SantaLucia J, Jr. (1998) Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. Biochemistry 37: 2170-2179.

26. Allawi HT, SantaLucia J, Jr. (1997) Thermodynamics and NMR of internal G.T mismatches in DNA. Biochemistry 36: 10581-10594.

27. Allawi HT, SantaLucia J, Jr. (1998) Thermodynamics of internal C.T mismatches in DNA. Nucleic Acids Res 26: 2694-2701.

28. Xia T, SantaLucia J, Jr., Burkard ME, Kierzek R, Schroeder SJ, et al. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 37: 14719-14735.

29. Papanicolaou C, Gouy M, Ninio J (1984) An energy model that predicts the correct folding of both the tRNA and the 5S RNA molecules. Nucleic Acids Res 12: 31-44.

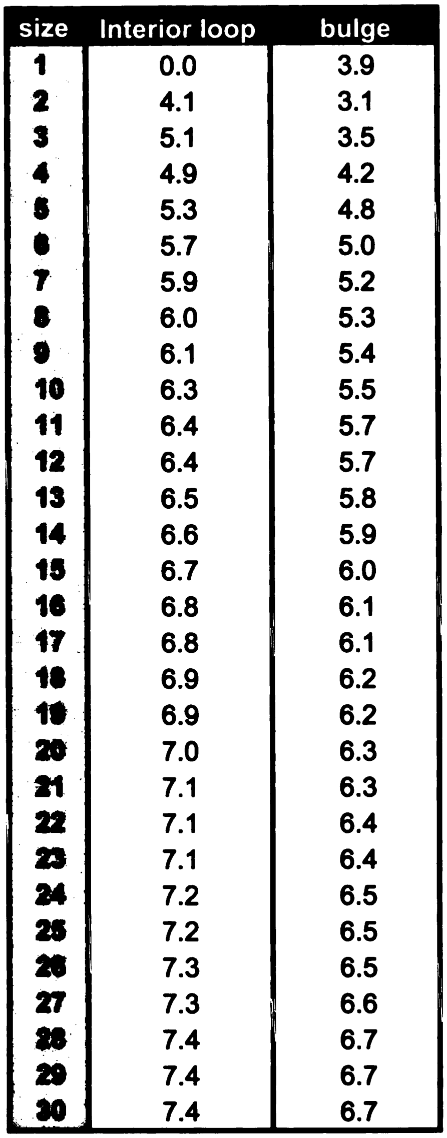
30. Peritz AE, Kierzek R, Sugimoto N, Turner DH (1991) Thermodynamic study of internal loops in oligoribonucleotides: symmetric loops are more stable than asymmetric loops. Biochemistry 30: 6428-6436.

31. Smith TF, Waterman MS, Fitch WM (1981) Comparative biosequence metrics. J Mol Evol 18: 38-46.

32. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31: 3406-3415.

33. Ziv J, Lempel A (1977) A universal algorithm for sequential data compression. IEEE Transactions on Information Theory 23: 337-343.

Table 2-1 Estimated internal loop destabilizing energy



size	Interior loop	bulge
1	0.0	3.9
2	4.1	3.1
3	5.1	3.5
4	4.9	4.2
5	5.3	4.8
6	5.7	5.0
7	5.9	5.2
8	6.0	5.3
9	6.1	5.4
10	6.3	5.5
11	6.4	5.7
12	6.4	5.7
13	6.5	5.8
14	6.6	5.9
15	6.7	6.0
16	6.8	6.1
17	6.8	6.1
18	6.9	6.2
19	6.9	6.2
20	7.0	6.3
21	7.1	6.3
22	7.1	6.4
23	7.1	6.4
24	7.2	6.5
25	7.2	6.5
26	7.3	6.5
27	7.3	6.6
28	7.4	6.7
29	7.4	6.7
30	7.4	6.7

The table is obtained from http://www.bioinfo.rpi.edu/~zukerm/cgi-bin/efiles.cgi?T=37#LOOP. The free energy parameters are for loop size equal or smaller than 30 bp and the unit is kcal/mol. For loops larger than 30 bp, an extra term, 1.75RTln(size/30), is added.

Part II. AOS Documentation

License

ArrayOligoSelector (AOS) is freely available under GPL license. Please acknowledge us approporiately if you use the program or any modules within the program. BLAT program is included in AOS releases as part of the components, which requires a license for commercial use. If you intend to use AOS for any commercial reason AND to use the "blat" or "gfclient" options, you need to obtain the approariate license for BLAT.

Download and installation

After successfully downloading AOS source code from the program's webpage (http://arrayoligosel.sourceforge.net), the code needs to be uncompressed first, and then the code is ready to use on a Linux system. For other UNIX systems, see section "System Requirements".

System requirements

Platform

It is easiest to set up AOS under the Linux operating system. AOS has been tested on Redhat Linux 6.1, 6.2, 8.0 and 9.0. It can be adapted to other UNIX environments such as Mac OS X. Users need to replace the following executables, blastall (NCBI), formatdb (NCBI), blat (UCSC) and gfclient (UCSC), according to the specific platform.

Mac OS X

Mac OS X users also need to have the Developers Tools package installed on the system.

Python

Python interpreter version 2.2 or above is required. It can be downloaded at http://www.python.org.

Input sequences

Input sequences and the complete genome sequences are needed. Both are required to be DNA sequences and in FASTA format.

Input sequences are the ORF sequences in the genome. The complete genome sequences should be either the complete set of gene sequences (exons), or the complete genomic sequences (exons, introns and intergenic regions). Two versions of AOS are provided for each scenario: exon version and contig version. Please refer to sections "Running AOS" and "Exon vs. Contig Version" for the differences in usage and implementation. If only partial genomic seqeunces are available, AOS will find unique oligos within the incomplete genome. Users should bear in mind that the oligos might have homologous sequences in the remaining partsof the genome.

Exon vs. Contig version - How does AOS define the cognate regions?

The cognate region of an ORF is the genomic location where the ORF originates. Accurate identification of cognate regions is very important because any homologous hits from these regions will be excluded in the uniqueness caculation.

The difference between the exon and contig versions lies in the conceptual definition and the identification of the cognate regions. In the exon version, the cognate region is simply defined as the input sequence itself. On the contrary, it is more complicated in the contig version. Each ORF is first aligned against the complete genomic sequence using BLAST or BLAT (user choice) , segments of 100% identical alignment are then stitched back together through a heuristic combinatorial process to recapitulate the exon pattern . Only combinations that can generate the original input sequence in a correct order are defined as the cognate regions. To accommodate cases where gene duplication exists, multiple cognate regions on different chromosomes are allowed. The cognate regions are recorded in the file "groupfile" in AOS's root directory. See section "Output File Description" for detailed description.

Running AOS

AOS has two sub-programs that run in series, the computation program followed by the selection program. The computation program calculates the following parameters for oligos starting from every position in the input sequence: uniqueness, sequence complexity, secondary structure, GC content and base pair starting position. Parameters generated by the computation program are stored in a series of output files, which are then used by the selection program to select an optimum set of oligos that are unique in the genome, with a low level of internal repeats and secondary structure, sharing a narrow range of GC percentage and close to the 3' end of the gene.

First Program – the computation program

Pick70_script1 (exon version) and Pick70_script1_contig (contig version) are the command line scripts to execute the computation program.

To run the scripts, users type "./Pick70_script1" or "./Pick70_script1_contig" on the command line and the usage instruction will show as screen output. Four command line arguments are required for Pick70_script1: filenames of the input and the genome sequence files, the length of the oligo and the exclusion of lowercase sequences. The exclusion argument has two choices: yes and no. If it is set as yes, then when an oligo has greater than 10% of its sequence in lowercase, parameters for the oligo will not be computed and the oligo will not be selected. Those oligos will be flagged with "F" in the computation program output files. This feature can be used in combination with a popular repeat masking program, Repeatmasker, to exclude highly repetitive sequences (such as low complexity regions and the alu element in the human genome) from computation and selection, thus speeding up the AOS computation program dramatically.

An additional argument is required for Pick70_script1_contig (five total). The fifth argument is the method for sequence alignment for cognate sequence identification and the choices are BLAST, BLAT or "gfclient". Both BLAT and gfclient are Blast-like alignment tools ideal for fast aligning exons to the genomes (Kent 2002). AOS runs faster if Blat or gfclient is used. Although BLAT and gfclient are essentially the same, gfclient requires setting up the gfserver in advance and BLAT uses more memory.

Two test files are provided: "test_input" and "test_genome".

Usage examples:

```
./Pick70_script1 test_input test_genome 70

./Pick70_script1_contig test_input test_genome no blat
```

```
./Pick70_script1_contig test_input test_genome no blast

./Pick70_script1_contig test_input test_genome yes blat

./Pick70_script1_contig test_input test_genome yes blast
```

Second program – the selection program

Pick70_script2 is the command line script to execute the selection program, which should be invoked after the computation program finishes.

To run selection program, users type "./Pick70_script2" on the command line and the usage instructions will show on the screen. Three command line arguments are required: the target GC percentage, the length of oligos, and the number of oligos per input sequence. There are also four optional arguments. The first optional argument is the user defined uniqueness cutoff as calculated in binding energy (the default value is top 5% and within 5 kcal/mol from the best uniqueness score in a given input sequence). The remaining three optional arguments are the nucleotide composition, maximum length and maximum tolerance level of user-defined exclusion (masking) sequence. These three optional arguments belong to a set, which should be provided together or neither should be provided. These masking arguments can be used to exclude stretches of sequences with only certain nucleotide compositions such as a long stretch of AT sequence. For example, the following argument combination "AT, 20, 0.1" represents the exclusion of any oligo with a continuous stretch of sequence that is longer than 20 bp with less than 10% G and C.

Usage Examples:

```
./Pick70_script2 28 70 1

./Pick70_script2 28 70 1 -35 20 AT 0.1
```

```
./Pick70_script2 28 70 1 -35 20 AT 0
```

Design parameters can be specified by users

- Target average GC content
- Number of oligos per input sequence
- Uniqueness cutoff (optional; if not specified, use default)
- User-defined masking sequence composition, length, and tolerance level (optional)

What are the ranges for the parameters?

In order to help users gain a better feel of the parameters of the computation program calculated for each oligo, we show here their ranges for 70mer oligonucleotides belonging to the test sequences that are included in the software distribution. Note these ranges will change for oligos of different length or generated from different source sequence.

- Uniqueness score (binding energy): 0 to –150 kacl/mol; smaller absolute value means a better oligo which is more unique and has less cross-hybridization.
- Secondary structure: 19 to 187; smaller value indicates a better oligo which has less secondary structures.
- Sequence complexity: 27 to 58; smaller value means a better oligo which has less low complexity sequences.
- GC content: GC percentage in the oligo sequence. The range is from 0 to 100 percent.

Output files

Final results files – generated by the selection program

- **oligo_fasta**

  "oligo_fasta" has the final selection results of the selected oligonucleotide probe sequences in FASTA format. The identifier of an oligo is its parent input sequence identifier plus the starting position of the oligo concatenated by the underscore character.

- **oligo_dup**

  "oligo_dup" stores the parameters for every oligo in the final selection. The parameters are scores for an oligo's GC percentage, secondary structure, sequence complexity, uniqueness and its genome targets,

  "oligo_dup" has the following format:

  Line 1 >oligo_id GC_percentage sequence_complexity_score secondary_structure_score

  Line 2 primary_target_identifier binding_energy_of_the_primary_target location_of_the_primary_target secondary_target_identifier binding_energy_of_the_secondary_target location_of_the_secondary_target

- **nodesign**

  "nodesign" stores the identifiers of input sequences which do not have any oligo selected because no oligo in those sequences can pass all the selection filters

(uniqueness, user-defined exclusion sequence composition, and not flagged by the lowercase option).

Intermediate results files – generated by the computation program

- **output0, 1, 2, ...**

  Oligo parameters calculated by the computation program are stored as a series of files named "output0", "output1" and ... , depending on the size of the results. Those files can be used by users who are interested in extracting the oligo parameters for other purposes or writing costomized selection program.

  The format for the "output" files is shown as the following. Each line records information for a single oligo. Fields are deliminated by the "TAB" character:

  Line 1 oligo_start_position uniquness_score GC_percentage
  sequence_complexity secondary_structure oligo_sequence
  primary_target_identifier primary_target_binding_energy
  primary_target_location secondary_target_identifier
  secondary_target_binding_energy secondary_target_location

- **groupfiles**

  The "groupfile" stores the information of the cognate genomic region for the input sequences. The file format is shown as the following:

  Line1 [input_sequence_identifier]

  Line 2 genomic_target_identifier

Line 3 +/- input_start_position input_end_position target_start_position target_end_position

"+/-" represents the plus or minus strand of the cognate region in its genomic target. If an input sequence is from a multi-exon ORF, lines 2 and 3 are repeated for each exon.

Speed

It took12 hours to design the 70mer oligonucleotide genome array for *Plasmodium falciparum* (12Mbp coding sequence, 23Mbp genomic sequence) on a dual 700MHz Linux system. The option used for cognate sequence identification was BLAT.

Chapter 3 . Application of the AOS system to the genome of *Plasmodium falciparum*

This chapter is a reprint from the following reference:

Zbynek Bozdech, Jingchun Zhu, Marcin P Joachimiak, Fred E Cohen, Brian Pulliam and Joseph L DeRisi. (2003) Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biology*, 4(2):R9.

ABSTRACT

Background

The worldwide persistence of drug-resistant *Plasmodium falciparum*, the most lethal variety of human malaria, is a global health concern. The *P. falciparum* sequencing project has brought new opportunities for identifying molecular targets for antimalarial drug and vaccine development.

Results

We developed a software package, ArrayOligoSelector, to design an open reading frame (ORF)-specific DNA microarray using the publicly available *P. falciparum* genome sequence. Each gene was represented by one or more long 70 mer oligonucleotides selected on the basis of uniqueness within the genome, exclusion of low-complexity sequence, balanced base composition and proximity to the 3' end. A first-generation microarray representing approximately 6,000 ORFs of the *P. falciparum* genome was constructed. Array performance was evaluated through the use of control oligonucleotide sets with increasing levels of introduced mutations, as well as traditional northern blotting. Using this array, we extensively characterized the gene-expression profile of the intraerythrocytic trophozoite and schizont stages of *P. falciparum*. The results revealed extensive transcriptional regulation of genes specialized for processes specific to these two stages.

Conclusions

DNA microarrays based on long oligonucleotides are powerful tools for the functional annotation and exploration of the *P. falciparum* genome. Expression profiling of

Background

Plasmodium falciparum, a parasitic protozoan, is the causative agent of the most lethal form of human malaria. It is responsible for 300-500 million infections per year in some 90 countries and regions throughout the tropical and subtropical world. Of these clinical cases, approximately 2.1 million result in death annually [1]. In areas where mosquito abatement has failed, chemotherapy, consisting of a limited selection of antimalarial agents, is the only defense against this disease. The increase in drug resistance throughout the malaria endemic regions is cause for great concern and calls for the development of new antimalarial measures, which would involve a larger variety of drug targets as well as a wider array of vaccine strategies (reviewed in [2,3]).

The study of malaria will be greatly helped by the publicly available complete genome sequence of *P. falciparum*. The sequencing project, driven by the Sanger Centre, the Institute for Genomic Research (TIGR), and Stanford University is essentially complete [4]. The sequence of the completed chromosomes are available for download from each sequencing center and from the *Plasmodium* Genome Resource, PlasmoDB [5,6]. Preliminary analysis of the 23 megabase-pair (Mbp) *P. falciparum* genome indicates the presence of approximately 5,400 genes spread across 14 chromosomes, a circular plastid genome and a mitochondrial genome. Strikingly, more than 60% of the predicted open reading frames (ORFs) lack orthologs in other genomes [4]. This fact underscores the need to elucidate gene function, yet many of the tools that have propelled the study of model organisms remain inefficient or nonexistent in *Plasmodium*. Despite recent improvements in *P. falciparum* transformation techniques, [7] the efficiency of stable transfection under a direct drug selection remains approximately 10^{-6}, making

knockout and gene replacement experiments difficult, and genetic complementation strategies nearly impossible. Genome-wide expression profiling by microarray technology provides an easy alternative for the functional genomic exploration of *P. falciparum*.

In organisms ranging from bacteria to humans, expression profiling has proved a powerful tool. Profiling has been used to gain important insights into processes such as development, responses to environmental perturbations, gene mutation, pathogen and host response, and cancer [8,9,10,11,12,13,14]. Expression profiling has already been successfully applied to the partial genome sequence of *P. falciparum*, and has been used to characterize the role of previously unannotated genes [15,16,17].

Here we present the design and assembly of a long-oligonucleotide *P. falciparum* gene-specific microarray using the currently available genomic sequence generated by the Malaria Genome Consortium [18,19,20]. During the course of this work, we have developed software, improved by experimental data and an open-source policy, for rapidly selecting unique sequences from predicted ORFs of any genome. Subsequently, we constructed a long-oligonucleotide-based *P. falciparum* microarray, which we used to evaluate changes in the global expression profile between two distinct stages of *P. falciparum* erythrocytic-stage asexual development - mid-trophozoite and mid-schizont. The large number of differentially expressed genes detected in this analysis suggests that extensive transcriptional regulation has a major role in the functional specialization of parasite development.

Results and discussion

P. falciparum ORF predictions

At the outset of these studies, a total of 27.6 Mbp of *P. falciparum* genomic sequence was obtained from the publicly available sources presented by the Malaria Genome Consortium [18,19,20] in October 2000. The sequence comprised two completely assembled chromosomes, the complete mitochondrial and plastid genomes, and the sum of all the partial contigs from the remaining chromosomes. ORF predictions were carried out using GlimmerM, a gene-finding tool trained with *P. falciparum* specific sequences [21,22]. Using default parameters, GlimmerM frequently yielded a large number of overlapping predictions (competing gene models) and thus additional filtering of the initial prediction output was required. As slight overprediction of ORFs is generally desirable for the purpose of expression array building, the post-prediction filtering of the GlimmerM output was modified with respect to the process used by the Malaria Genome Consortium [21]. Briefly, individual predictions that overlapped and were on opposite strands or in different reading frames were retained. For competing predictions within a given GlimmerM gene model, ORFs that were extended downstream by at least 300 bp and were within 300 bp of the total size compared to the size of the largest prediction were chosen. In all other cases, the largest predicted ORF was selected. This selection method resulted in 290 ORF predictions for chromosome 2, whereas the Malaria Genome Consortium selected 210 for the same chromosome [21].

The first round of predictions, carried out on the publicly available genomic sequence as of August 2000, yielded 8,008 putative ORFs. The predicted ORFs are available as additional data with the online version of this paper (see Additional data files) and from [23]. As a first step to annotation, the translation of all predicted ORFs

were used to search the Astral, SwissProt, and non-redundant (NR) databases for sequence similarities using the Smith-Waterman algorithm [24]. In addition, all ORF predictions were linked to their counterparts in PlasmoDB [5,6].

ArrayOligoSelector: array element design

To construct a gene-specific microarray of the *P. falciparum* genome, we designed 70 mer oligonucleotide array elements. We chose this length for a number of reasons. Long oligonucleotides are a highly sensitive alternative to PCR products and provide a means to readily distinguish between genes with high degrees of sequence similarity [25]. In addition, the presence of various types of repetitive sequences and highly homologous gene families in the AT-rich *P. falciparum* genome contributes to a high rate of PCR failure ([17] and J.L.D., unpublished results). A software program, ArrayOligoSelector, was developed specifically for the purpose of systematically selecting gene-specific long oligonucleotide probes for entire genomes. The latest version and complete source code for ArrayOligoSelector is freely available at [26]. For each ORF, the program optimizes the oligonucleotide selection on the basis of several parameters, including uniqueness in the genome, sequence complexity, lack of self-binding, and GC content (Figure 1). Similar approaches to oligonucleotide design have previously been described, but the exact algorithms, source code, and/or accompanying hybridization data are not available [25,27,28].

ArrayOligoSelector helps ensure complete genome coverage and optimal array hybridization while avoiding several potential problems originating from the peculiar characteristics of the *P. falciparum* genome. The algorithm attempts to minimize cross-

hybridization between the oligonucleotide and other regions of the genome. To evaluate the potential for cross-hybridization, early versions of ArrayOligoSelector used a simple BLASTN alignment identity [29]. Although this method prevents the selection of troublesome sequences, it does not take into account the effect of mismatch distribution or base composition. Subsequent versions of ArrayOligoSelector were improved by calculating a theoretical energy of binding between the oligonucleotide and its most probable cross-hybridization target in the genome ('second best target'). The binding energy (kcal/mol) is calculated using a nearest-neighbor model using established thermodynamic parameters [30,31,32,33,34,35]. Thus, a sequence with high cross-hybridization potential will have a more stable binding energy with a larger absolute value. In contrast, a sequence unique in the genome will yield a smaller absolute value for the binding energy. A representative plot of the calculated binding energies for all possible 70 bp oligonucleotides from a putative *var* gene (PlasmoDB v4.0 annotated gene ID PF08_0140) is shown in Figure 2a.

An important aspect of oligonucleotide design for microarray hybridization is avoiding secondary structures within the oligonucleotide, as these are likely to be detrimental to hybridization performance. To avoid selecting oligonucleotides with secondary complex structure, ArrayOligoSelector uses the Smith-Waterman algorithm with the PAM47 DNA matrix to calculate the optimal alignment score between the candidate oligonucleotide sequence and the reverse complement of that sequence [24]. A high Smith-Waterman score indicates the potential to create secondary structures (Figure 2b).

The presence of low-complexity sequence could also result in significant nonspecific cross-hybridization. For example, the *P. falciparum* genome contains a large number of low-complexity sequence elements as a result of a high frequency of continuous stretches of A and T nucleotides in both the non-coding and the coding regions. ArrayOligoSelector automatically detects such sequences by subjecting candidate oligonucleotide sequences to a lossless compression [36]. The compression score, calculated as the difference in bytes between the original sequence and the compressed version, is inversely proportional to complexity (Figure 2c). Using this score, repeats of essentially any nature are detected in a computationally efficient manner.

In addition, in order to avoid specific sequence features, ArrayOligoSelector supports filtering based on user-defined patterns. This feature can be used to implement filtering rules based on empirically derived data. Finally, the melting temperature of an oligonucleotide is largely determined by its GC content. As is the case with most ORFs, there exists a large range of %GC values (< 10 to > 60%) over a 70 bp window (Figure 2d). For this reason, a user-defined %GC target range is used by ArrayOligoSelector such that the majority of the array elements will share a similar base composition and hybridization properties across the array.

Given the above parameters, ArrayOligoSelector evaluates every 70 mer sequence within an ORF and chooses an optimal set on the following criteria. The uniqueness-filter requires oligonucleotides to satisfy two simultaneous threshold criteria based on the calculation of the binding energy to their second-best target (the best target is itself). First, the oligonucleotide must rank among the top 5% of the unique or almost unique 70 mers in the entire ORF. Second, its binding energy must be within 5 kcal/mol of the best

candidate for the ORF. In addition, an optional user-defined energy threshold can operate in conjunction with the default threshold. Initial settings for the low-complexity and the self-binding terms allow the top-scoring 33% of 70 mers to pass to the next selection step. Finally, an optional 'user-defined sequence filter' simply eliminates the 70 mer candidates containing the defined sequences. These four filters operate on the entire set of 70 mer candidates for a particular ORF and generate four independent output sets. The intersection of the four outputs is then subjected to the final selection. If no common oligonucleotide is identified in the first intersection, the self-binding and complexity filters are incrementally relaxed until an intersection becomes available. The final selection of candidate oligonucleotides depends upon the %GC filter and 3'-end proximity ranking. Initially, oligonucleotides are allowed to pass if they meet the user-specified %GC. If no oligonucleotide with the desired GC content is found, the target %GC range is relaxed by one percentage point in each direction until one or more oligonucleotides pass. As a final step, a single candidate, closest to the 3' end of the gene is chosen. Finally, ArrayOligoSelector generates an output file containing the oligonucleotide selections for each putative ORF.

From our initial set of predictions, a total of 6,272 70 mer oligonucleotides were selected and synthesized. For our first pass of malaria oligonucleotide selections, the earlier version of ArrayOligoSelector utilizing the BLASTN-based identity threshold was used. The identity cutoff was adjusted to a very conservative value of < 30 bp of identity. The initial setting of the GC content filter was set to 28% GC (73°C T_m. Subsequently, with the release of additional sequence information, a new set of predictions was generated in April 2002 and an additional 1,025 oligonucleotides were selected using the

upgraded version of ArrayOligoSelector. In this selection, the user-defined uniqueness threshold was set at -35 kcal/mol, the value at which cross-hybridization is essentially eliminated (Figure 3). The GC content target was set at 28%. The sequence and location of each oligonucleotide is available online [23]. The experiments described in the following section were conducted with the first set of predictions only. As additional annotations become available for the whole genome sequence, additional oligonucleotides will be selected and added to the existing collection. We expect the final set to contain approximately 8,500 oligonucleotides.

Oligonucleotide performance

Hughes *et al.* [25] showed that 60 mer oligonucleotides make highly sensitive specific microarray elements for expression profiling of *Saccharomyces cerevisiae* [25]. The oligonucleotides used in that study were synthesized *in situ* using ink-jet technology whereas the oligonucleotides used in our study were commercially synthesized and subsequently printed using mechanical deposition. Similarly to the experiments of Hughes *et al.*, we wished to test experimentally the effect of mismatches on sensitivity and specificity of 70 mer oligonucleotides in the context of a complex hybridization mixture (*P. falciparum* total RNA). Ten separate malaria ORF predictions were arbitrarily selected for analysis and for each of these ORFs a set of ten oligonucleotides were synthesized. The first oligonucleotide in each set represents the original 70 mer selection from ArrayOligoSelector. Each successive oligonucleotide within a set contains an increasing number of mutations made in increments of 10%. Thus, the second oligonucleotide in each set had seven bases (10%) altered, while the last oligonucleotide

had 63 bases (90%) mutated. For the first set of five ORFs (Figure 4a,4b,4c,4d,4e), which is referred to as the 'distributed set', both the position and the identity of each mutation was random. For the second set of five ORFs (Figure 4g,4h,4i,4j,4k), referred to as the 'anchored set', the mutations in each oligonucleotide were limited to the ends of the sequence. In this manner, a contiguous stretch of perfectly matched bases was always preserved in the center of each oligonucleotide.

Figure 4 summarizes normalized hybridization intensities of control oligonucleotides obtained from the global gene-expression comparisons between trophozoite and schizont stages. The results originate from the six microarray hybridizations presented in Figure 5 and 10 additional hybridizations available as additional data files [23]. The resulting hybridization intensity measurements for each oligonucleotide were averaged across all hybridizations and scaled as a fraction of the average intensity of the perfect-match oligonucleotide for each set. As is evident from Figure 4a,4b,4c,4d,4e, the presence of internal mismatches (bubbles and bulges) had a large effect on hybridization performance: oligonucleotides with 10% mismatches (7 bases) suffered an average reduction of 64% in hybridization intensity when compared to the perfect match, while oligonucleotides with 20% (14 bases) or more mismatches were reduced by an average of 97% (Figure 4f). For the anchored set (Figure 44g,4h,4i,4j,4k), a more gradual hybridization trend was observed. Mutating the terminal 14 bases (7 bases at each end) resulted in an average loss of 49% of the maximal hybridization intensity. Not until 42 bases had been mutated (21 bases at each end) did the relative intensity of hybridization drop by an average of 97.5% (Figure 4l). In agreement with the findings of Hughes *et al.* [25], the data from the anchored set of oligonucleotides reveal a strong

relationship between the length of contiguous match (the equivalent of oligonucleotide length) and overall hybridization performance.

To measure the extent to which the energy calculation implemented in ArrayOligoSelector matches reality, we have plotted in Figure 3 the calculated energy of the 100 control oligos shown in Figure 4 and their relative intensities of hybridization. The calculated energy and relative intensity of hybridization correlate at $|r| = 0.91$. For comparison, the relative intensity of hybridization and number of nucleotide identities correlate at $|r| = 0.72$. This indicates that a calculated binding energy approach may be used to estimate the potential for cross-hybridization for any sequence relative to the rest of the genome. The specificity for each oligonucleotide is uniquely and computationally determined and expressed as a binding energy (kcal/mol).

To further address the question of specificity of oligonucleotide hybridization to their targets in a complex sample we introduced a set of probes targeting a set of 19 non-repetitive sequences from *S. cerevisiae* to the microarray. To control for the nucleotide bias of the malaria genome relative to yeast, the selection criteria for this set were identical to selection of the plasmodial microarray elements. The average GC content of the *S. cerevisiae* oligonucleotides was 31.5%, whereas the average GC content of plasmodial oligonucleotides is 32.5%. The average signal-to-background ratios across all hybridizations for these negative control spots was less than twofold, which is well below the conservative fivefold signal-to-background threshold used to filter data (see Materials and methods). In addition, a series of 10 hybridizations was carried out where total RNA from an asynchronous parasite culture was hybridized against PCR products corresponding to the negative control *S. cerevisiae* sequences. In these hybridizations the

yeast PCR fragment hybridized strictly to its cognate sequence, while the average signal to background value for plasmodial elements in the same channel was 1.17 ± 0.05. In no individual case did a plasmodial element yield a signal greater than 2.3% of the target hybridization signal intensity. The results of these microarray hybridizations are available as additional data files [23].

To assess whether separate oligonucleotides designed to represent the same target gene perform in a similar manner, we examined three distinct situations: elements dispersed over a long single exon ORF (Figure 6a), overlapping oligonucleotides (Figure 6b), and oligonucleotides representing multiple exons of a single gene (Figure 6c). In each case we observed consistent oligonucleotide performance.

Gene-expression profiling of trophozoites and schizonts

We chose a direct comparison of the trophozoite and schizont stages of the *P. falciparum* asexual intraerythrocytic life cycle as a first step toward comprehensively profiling all life-cycle stages of this parasite. The trophozoite and schizont represent two distinct developmental stages within the 48-hour plasmodial erythrocytic life cycle. These stages vary greatly in morphology, biochemical properties, and transcriptional activity (reviewed in [15,37]). The mid-trophozoite stage, 18-24 hours post-invasion, contains a highly transcriptionally active nucleus with abundant euchromatin. In addition, trophozoites are characterized by massive hemoglobin ingestion, intake of nutrients from the surrounding medium, increasing concentration of cytoplasmic ribosomes and rapid formation of organelles. In contrast, the mid-schizonts, at 36-42 hours post-invasion, are characterized by DNA replication (16-32 copies) and compaction into newly formed

nuclei. In addition, maturation of merozoite cells begins at the schizont stage and is characterized by the appearance of merozoite organelles such as the rhoptry and dense granules. The several trophozoite- and schizont-specific genes identified previously provide an excellent source of positive controls for the experiments described below.

For microarray hybridization, total RNA was prepared from synchronized *in vitro* *P. falciparum* cultures representing the trophozoite stage and the schizont stage (see Materials and methods). Six independent hybridizations were carried out; in three, the trophozoite-derived cDNA was labeled with Cy3 and the schizont-derived cDNA with Cy5. In the other three hybridizations, the fluorophore assignment was reversed. Of the genes assayed, 854 features displayed a differential expression greater than twofold (Figure 5): 525 showed higher relative transcript abundance in trophozoites than in schizonts, whereas 326 had greater relative transcript abundance in schizonts. Linear regression ratios were calculated for each possible pair of microarray hybridizations using the filtered dataset. The correlation between hybridizations with the same Cy3/Cy5 assignment was $r = 0.94 \pm 0.02$, while correlation of hybridizations with the opposite Cy3/Cy5 order was $r = 0.89 \pm 0.03$.

Northern blot hybridizations

To confirm the microarray results, we examined six genes by northern blot analysis. In the microarray hybridization, the expression levels of two of the selected genes were unchanged (< 2-fold) while four additional genes showed a differential expression between the trophozoite and schizont stage (> 2-fold). An equal mass of total RNA from both the trophozoite and schizont stages was hybridized with PCR-generated

DNA probes corresponding to the selected genes. Subsequently, each northern blot was stripped and rehybridized with a probe specific for the 41 kD antigen (p41), fructose-bisphosphate aldolase (PfALDO; PlasmoDB v4.0 ID PF14_0425; Oligo ID M11919_1) [38], as a loading control. While the relative amount of PfALDO transcript differs by more than twofold between trophozoite and schizont stages when equal masses of total RNA are blotted, we found that the relative amount of PfALDO to be essentially equivalent when equal masses of poly(A)$^+$ RNA were used for the northern blot (Figure 7a). The discrepancy between northern blots with total RNA and poly(A)$^+$ mRNA are probably due to changes in the relative amounts of mRNA and ribosomal RNA during the intraerythrocytic life cycle. The poly(A)$^+$ northern blot measurements agree well with the replicate array hybridizations, in which PfALDO was consistently less than 1.5-fold differentially expressed (Figure 7b). To make northern blot measurements comparable to the normalized expression array ratios, the ratio between the two stages was measured using a phosphoimager and divided by the ratio obtained for the PfALDO control in each case. The normalized ratios of the radiolabel signal were highly consistent with the averaged ratios from the six microarray hybridizations (Figure 7b).

Biological significance of the gene-expression results

The genome-wide expression data summarized by hierarchical cluster analysis (Figure 5) resulted in two main gene categories, corresponding to genes differentially expressed between trophozoite and schizont stages. Serving as internal positive controls, a number of previously well-characterized plasmodial genes were detected in both categories. In addition, an evaluation of the homology-based gene identities within these

categories revealed several functional gene groups. All data from these experiments are available at PlasmoDB and the DeRisi Lab website [23].

Trophozoite genes

The predominant group of features with elevated expression in the trophozoite stage comprised genes encoding various components of the eukaryotic translation machinery. This group contained 24 and 33 proteins of the 40S and 60S ribosomal subunits (RPS and RPL), respectively. In addition, nine orthologs of aminoacyl-tRNA synthetases, and 10 initiation and seven elongation translation factors were detected among trophozoite-specific genes. Several previously identified plasmodial genes were present in this group, including Asp-tRNA synthetase, two plasmodial elongation factors (PfEF1A and PfEF2) and one ribosome-releasing factor, PfRF1 [21,39]. Consistent with our findings, PfEF-1A has been previously shown to have peak expression during the trophozoite stage [40]. Two additional gene groups whose functions are linked to the process of protein synthesis were present among the trophozoite genes: five DEAD-box RNA helicases, including a close homolog of *P. cynomolgi* RNA helicases-1 [41] and 23 molecular chaperone-like molecules, including two *P. falciparum* heat-shock proteins such as PfHSP70 (GenBank accession number M19753) and PfHSP86 (accession number L34028), and a homolog of a DnaJ-domain-containing protein family, DNJ1/SIS1 homolog [42]. These data agree with previous studies that found a group of DEAD-box RNA helicases to be overexpressed during the trophozoite stage in *P. cynomolgi* [41]. Along with the genes for the translation machinery a number of genes involved in various steps of RNA synthesis and processing were located among the 'trophozoite genes',

including 16 ORFs belonging to various RNA polymerase complexes and 11 splicing factors (Figure 5). Two previously identified plasmodial RNA polymerase components were found in this group, including the largest subunit of *P. falciparum* RNA polymerase II, PfRNApolIIA (M73770), and a homolog RNApolK (14 kD) [39]. The expression characteristics revealed are also consistent with several previous studies that suggested that the plasmodial transcription and translation machinery is active through the late ring and early trophozoite stage before decaying during the late schizont stage [15,40].

Another functional group of genes that encode enzymes of cellular biosynthetic pathways was distinguished within the trophozoite category. This gene set includes 16 enzymes of carbohydrate metabolism, 10 ORFs likely to be involved in nucleotide metabolism, and 11 ORFs involved in the biosynthetic pathways of several amino acids. Several well-characterized plasmodial genes were identified in this metabolic collection, including *P. falciparum* lactate dehydrogenase (PfLDH; 027743), enolase (U00152), triose-phosphate isomerase (PfTPI; L01654), glucose-6-phosphate isomerase (PfG6PI; J05544), hypoxanthine-guanine phosphoribosyl-transferase (PfHGPRT; X16279) and dihydropteroate synthetase (PfDHPS; U07706). In addition, a group of 11 proteolytic enzymes potentially involved in hemoglobin degradation was detected among the trophozoite genes; these include a cysteine protease, falcipain-2 (AF251193), a metalloprotease falcilysine, (AF123458), and a member of an aspartic protease family, plasmepsin-2 (L10740). Falcipain-2 and plasmepsin-2 have been the targets of recent drug discovery research [43,44].

Overall, the emergent gene clusters suggest that the trophozoite stage, a central phase of plasmodial intraerythrocytic development, is characterized by the activation of

general cellular growth functions such as transcription, translation and hemoglobin degradation and biosynthesis of basic metabolites.

Schizont genes

A large number of ORFs found in the schizont-expressed category correspond to genes previously associated with the various steps by which newly released merozoites invade new host cells. The initial step of this process, adhesion of the merozoite to the surface of an erythrocyte, is facilitated by several classes of proteins exposed on the surface of the parasite. Eighteen ORFs, identical or homologous to proteins associated with the merozoite surface, were present among the schizont-enriched genes. This group included four merozoite surface proteins (MSP): MSP1 (M19753), MSP4, MSP5, (AF033037) and MSP6 (AY007721). Additional members of this group include two ORFs containing Duffy-like binding domains, erythrocyte-binding antigen, EBA 175 (L07755), a putative erythrocyte-binding protein, EBL1 (AF131999), and proteins known to be delivered to the surface from apical organelles, including apical membrane antigen, AMA1 (U65407), and finally two rhoptry-associated proteins (RAP1 (U20985) and RAP2).

Initial attachment of the merozoite is followed by reorientation of the parasite cell with its apical part toward the erythrocyte membrane followed by invagination of the membrane. Previous studies suggested that both steps are facilitated by the action of actomyosin, which requires ATP hydrolysis [45]. Consistent with these findings, we found five proteins previously associated with this process, *pf-actinI* (M19146), *pf-myoA* (AF255909), and merozoite cap protein-1 (U14189), and two subtilisin-like proteases

(PfSUB1 and PfSUB2 (AJ132422)) differentially enriched in schizonts. Interestingly, one additional homolog of PfSUB1 was identified among the schizont genes. Moreover, the expression levels of a set of plasmodial protein kinases were previously found to be augmented during the late stages of the malarial erythrocytic life cycle [15]. Our findings confirm and extend this report: 26 unique ORFs sharing a high to medium level of homology with protein kinases and phosphorylases had elevated mRNA levels during the schizont stage (Figure 5). Two previously identified representatives were present in this set: a cAMP-dependent protein kinase, PfPKAc (AF126719), and a plasmodial serine/threonine protein phosphatase, PfPPJ (AF126719).

A second functional group of genes with increased expression in schizonts encodes proteins that are thought to function on the periphery of a newly infected erythrocyte at the early stages of asexual development. Representatives include: the genes for ring-infected erythrocyte surface antigen (RESA) (X04572) and several close RESA homologs, CLAG9 (AF055476) the related gene *CLAG3.1*, and two members of the serine-repeat rich protein (SERA) family [21]. In addition to these well-characterized surface proteins, the schizont-enriched set of transcripts contained a number of ORFs identical or homologous to proteins recognized by antibodies present in plasmodium immune sera obtained either from model organisms [46] or from acute and/or convalescent patients [47]. In summary, the schizont stage of plasmodial development featured genes predominantly occupied with the process of merozoite function as well as the advance synthesis of transcripts for proteins that facilitate parasite establishment within the newly infected erythrocyte.

Taken together, these results suggest that the parasite cell in the trophozoite stage is dedicated to cell growth, and the predominant function of the mid-late schizont stage is maturation of the next generation of merozoites. Of particular interest is the large number of ORFs within both categories (39% in trophozoite and 61% in schizonts) with no putative functions assigned. These ORFs have little to no homology to any other known genes and may possibly represent highly specialized functions not likely to be shared outside this family of parasites.

Conclusion

In this study, we present a *P. falciparum* ORF-specific microarray utilizing 70 mer oligonucleotides as individual microarray elements. This approach helped to overcome potential problems originating from low PCR amplification and allowed us to select probes with a high specificity, thereby minimizing potential cross-hybridization. Moreover, the oligonucleotide-selection algorithm allowed a balanced GC content (around 28%) across the entire microarray set, which is significantly higher than the plasmodial genome average, which is 19.4% with 23.7% in coding regions [4].

Application of the ArrayOligoSelector is not restricted to the *P. falciparum* genome, but is broadly useful for the automated selection of hybridization probes for a range of species. The flexibility of the selection parameters controlling stringency of uniqueness, self-binding, complexity, user-defined filters and GC content, allows the selection of oligonucleotides appropriate for any genome.

Evaluation of results from derivative control oligonucleotides showed that long oligonucleotides could tolerate 10% mismatches; however, alteration of the target

sequence by more then 20% eliminated most of the hybridization signal. Therefore, small sequencing errors and natural variation among isolates are not likely to impact on sensitivity. These performance characteristics imply that the array design for this effort can accommodate the study of essentially any *P. falciparum* strain with a high degree of specificity.

At present, the *P. falciparum* microarray used in this study consists of approximately 6,000 gene-specific elements corresponding to the majority of the total coding content predicted for the *P. falciparum* genome. As new sequence and improved gene predictions arise, additional elements will be added to this evolving platform. Moreover, the present oligonucleotide representation could be further extended for investigation of several unusual *P. falciparum* genetic and transcriptional phenomena, including antisense mRNA transcription [48] and alternative splicing and/or transcriptional initiation [49,50]. This may be achieved by designing exon-specific array features, as well as antisense oligonucleotides. The oligonucleotide collection could also be expanded by sequences corresponding to intergenic genomic regions. Inclusion of such elements was found to be extremely useful for identifying protein-binding DNA regions by chromatin-immunoprecipitation as well as genes not detected by automated gene-prediction algorithms [51].

Within both the trophozoite and schizont categories, large numbers of genes belong to functionally related processes. These include genes encoding ribosomal subunits, multiple factors for transcription and translation, enzymes of biosynthetic and catabolic pathways, or merozoite adherence and invasion machinery. These results are consistent with predictions that a large number of plasmodial genes undergo strict stage-

specific transcriptional regulation, and that such (co-)regulation is shared among functionally related genes [15,52]. Naturally, a 'fine-resolution' global gene-expression profile including the different steps of the plasmodial life cycle for multiple divergent strains will be necessary to characterize fully the intraerythrocytic life of the parasite. At present, our laboratory is analyzing a global gene-expression profile of the 48-hour erythrocytic life cycle with 1-hour resolution for three strains of *P. falciparum*.

In a number of model organisms, high-resolution gene-expression maps have served as extremely powerful tools for discovery and characterization of novel genes as well as exploration of multiple cellular functions [9,11]. The gene-expression maps typically comprise genome-wide expression profiles at a number of different stages of cellular development, profiles of multiple strains and genetic variants, and global expression responses to number of growth perturbations and growth-inhibitory drugs. Following a similar approach in *P. falciparum* is most likely to provide substantial information about the many ORFs that lack functional annotation. Further understanding of cellular physiology of this parasite including basic metabolic functions and the intricate interactions between the parasite cell and human host immune system will be a key step in uncovering new targets for antimalarial drug discoveries and vaccine development.

Materials and methods

Microarray fabrication

The 70-bp oligonucleotides were synthesized (Operon Technologies, CA), resuspended in 3 × SSC to a final concentration of 60 pmol/μl, and spotted onto poly-L-

lysine-coated microscopic slides, as previously described [53]. All oligo sequences are available at [23].

Cell cultures

P. falciparum parasite cells (W2 strain) were cultured as described [54] with slight modifications: 2% suspension of purified human red blood cells in RPMI1640 media supplemented with 0.25% AlbumaxI (GIBCO/Invitrogen, San Diego, CA), 2 g/l sodium bicarbonate, 0.1 mM hypoxanthine, 25 mM HEPES pH 7.4, and 50 μg/I gentamycin. Cells were synchronized by two consecutive sorbitol treatments on two consecutive cell cycles (a total of four treatments) and harvested at the subsequent trophozoite stage (18-24 h post-invasion) and schizont stage (36-42 h post-invasion). For the trophozoite stage collection, visual inspection of the Giemsa stains show a nearly pure trophozoite population with less than 1% schizonts. For the schizont stage collection, we estimate the amount of ring contamination to be around 3%. The cells were harvested in prewarmed PBS at 37°C, and spun at 1,500 *g* for 5 min. Cell pellets were rapidly frozen in liquid nitrogen and stored at -80°C.

RNA preparation and microarray hybridization

Total RNA was prepared directly from the frozen pellets of parasitized erythrocytes, where approximately 1 ml of cell pellet was lysed in 7.5 ml Trizol (GIBCO) and RNA was extracted according to the manufacturer's instructions. mRNA was isolated from total RNA preparations using the Oligotex mRNA Mini Kit (Qiagen, Valencia, CA). For the hybridization experiments, 12 μg total RNA was used for first-strand cDNA

synthesis as follows: RNA was mixed with a mixture of random hexamer (pdN$_6$) oligonucleotides and oligo-(dT$_{20}$) at final concentration 125 μg/μl for each oligonucleotide. The mixture was heated to 70°C for 10 min and then incubated on ice for 10 min. Reverse transcription was started by adding dNTPs to a final concentration of 1 mM dATP and 500 μM each: dCTP, dGTP, dTTP and 5-(3-aminoallyl)-2'-deoxyuridine-5'-triophosphate, (aa-dUTP) (Sigma), with 150 units of StrataScript (Stratagene, La Jolla, CA). The reaction was carried out at 42°C for 120 min and the residual RNA was hydrolyzed with 0.1 mM EDTA and 0.2 M NaOH at 65°C for 15 min. The resulting aa-dUTP-containing cDNA was coupled to CyScribe Cy3 or Cy5 (Amersham, Piscataway, NJ) monofunctional dye in the presence of 0.1 M NaHCO$_3$ pH 9.0. Coupling reactions were incubated for a minimum of 1 h at room temperature. The labeled product was purified using QIAquick PCR purification system (Qiagen). Hybridizations and final washing procedures were carried out as described [9] with slight modifications. Briefly, the hybridization medium contained 3 × SSC, 1.5 μg/μl poly(A) DNA (Pharmacia Biotech, Uppsala), and 0.5% SDS. Hybridizations were incubated at 65°C for 8-16 h. Arrays were washed in 2 × SSC/0.2% SDS and then 0.1 × SSC at room temperature. The microarrays were scanned with a GenePix 4000B scanner and the images analyzed using GenePix Pro 3.0 software (Axon Instruments, Union City, CA). Subsequently, the data were normalized using the AMAD microarray database and subjected to the cluster analysis using the CLUSTER and TREEVIEW software, as described [53]. For the CLUSTER analysis, low-quality features and features with a signal level less than fivefold the background were filtered from the initial raw data set, yielding 4,737 elements. Subsequently, features with an arbitrary twofold fluorescence

signal difference in at least four experiments were considered. All programs and microarray-related protocols are available online [55].

Probe preparation and northern blot analysis

The northern blot probes were generated by PCR using the following oligonucleotide sequences:

FWD-M11919_1:	5'-TAGAAAACAGAGCTAGCTACAGAG;
REV-M11919_1:	5'-AGTTGGTTTTCCTTTGGCTGTGTG;
FWD-M1282_7:	5'-CTGTAGGTGGTATCCCTTTACAAG;
REV-M12812_7:	5'-GACAAATAATAATGCCATACCAGG;
FWD-I12861_2:	5'-AAATGCAGTTGTTACTGTCCCTG;
REV-I12861_2:	5'-GCTCTTTTGTCAGTTCTTAAATCG;
FWD-F5910_2:	5'-ACAACCAGTTTGCTCTGCTTATC;
REV-F5910_2:	5'-GGCCGACATTAATTGCTTATATGC;
FWD-M38757_7:	5'-TAGAAGTATATCATTCCGAAGGTG;
REV-M38757_7:	5'-GTAGAAGCTTCAATATCAAGCTC;
FWD-M1282_7:	5'-CTGTAGGTGGTATCCCTTTACAAG;
REV-M12812_7:	5'-GCTAATGCCTTCATTCTCTTAGTT;
FWD-Ks44_1:	5'-GGCAAGCTATAACAAATCCTGAGA;
REV-Ks44_1:	5'-GCTAAAGCGGCAGCAGTTGGTTCA.

Total RNA (10 µg) or poly(A)$^+$ RNA (0.4 µg) was resolved on a denaturing 1% agarose gel, transferred to nitrocellulose membrane and hybridized with a radiolabeled

probe as described [56]. The blots were analyzed using ImageQuant v1.2 (Molecular Dynamics, Sunnyvale, CA).

Additional data files

The predicted ORFs and GenePix results (GPR) files containing raw data for Figure 5 and from 10 additional hybridizations are available as additional data files with the online version of this paper and from [23]. Data for Figure 5: three hybridizations (1, 2,3) with trophozoite RNA labeled with Cy3 and schizont RNA labeled with Cy5; Three hybridizations (4,5,6) with trophozoite RNA labeled with Cy5 and schizont RNA labeled with Cy3. Additional hybridizations: Six hybridizations (7,8,9,10,11,12) with trophozoite RNA labeled with Cy3 and schizont RNA labeled with Cy5; four hybridizations (14,15,15,16) with trophozoite RNA labeled with Cy5 and schizont RNA labeled with Cy3.

ORF predictions of August 2000 were predicted from contig sequences available in August 2000, using GlimmerM software. These predictions were used to design the first set of 70 mer oligonucleotides and includes genes from the plastid genome. ORF predictions of October 2000 were predicted from contig sequences available in October 2000, using GlimmerM software. These predictions also include genes from the plastid and mitochondrial genomes.

Acknowledgements

We thank Pradip Rathod and his lab for ongoing discussion and assistance. We also thank the Howard Hughes Medical Institute, Anita Sil, Phil Rosenthal, David Roos

and Patrick O. Brown. Z.B., M.P.J., and J.L.D. are supported by an award from the Searle Foundation and the Burroughs Wellcome Fund.

References

1. Sachs J, Malaney P: The economic and social burden of malaria. *Nature* 2002, 415:680-685.

2. Ridley RG: Medical need, scientific opportunity and the drive for antimalarial drugs. *Nature* 2002, 415:686-693.

3. Richie TL, Saul A: Progress and challenges for malaria vaccines. *Nature* 2002, 415:694-701.

4. Gardner MJ, Hall N, Fung E, White O, Berrlman M, Hyman R, Carlton JM, Pain A, Nelson K, Bowman S, *et al.*: Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 2002, 419:498-511.

5. PlasmoDB: The Plasmodium Genome Resource [http://plasmodb.org]

6. Bahl A, Brunk B, Coppel RL, Crabtree J, Diskin SJ, Fraunholz MJ, Grant GR, Gupta D, Huestis RL, Kissinger JC, *et al.*: PlasmoDB: the *Plasmodium* genome resource. An integrated database providing tools for accessing, analyzing and mapping expression and sequence data (both finished and unfinished). *Nucleic Acids Res* 2002, 30:87-90.

7. Deitsch K, Driskill C, Wellems T: Transformation of malaria parasites by the spontaneous uptake and expression of DNA from human erythrocytes. *Nucleic Acids Res* 2001, 29:850-853.

8. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM: Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996, 14:457-460.

9. DeRisi JL, Iyer VR, Brown PO: Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997, 278:680-686.

10. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, *et al.*: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 2001, 98:10869-10874.

11. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, *et al.*: Functional discovery via a compendium of expression profiles. *Cell* 2000, 102:109-126.

12. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, *et al.:* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med* 1998, 4:1293-1301.

13. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 1998, 9:3273-3297.

14. de Avalos SV, Blader IJ, Fisher M, Boothroyd JC, Burleigh BA: Immediate/early response to *Trypanosoma cruzi* infection involves minimal modulation of host cell transcription. *J Biol Chem* 2002, 277:639-644.

15. Ben Mamoun C, Gluzman IY, Hott C, MacMillan SK, Amarakone AS, Anderson DL, Carlton JM, Dame JB, Chakrabarti D, Martin RK, *et al.*: Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite *Plasmodium falciparum* revealed by microarray analysis. *Mol Microbiol* 2001, 39:26-36.

16. Hayward RE: *Plasmodium falciparum* phosphoenolpyruvate carboxykinase is developmentally regulated in gametocytes. *Mol Biochem Parasitol* 2000, 107:227-240.

17. Hayward RE, Derisi JL, Alfadhli S, Kaslow DC, Brown PO, Rathod PK: Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. *Mol Microbiol* 2000, 35:6-14.

18. The Sanger Centre *Plasmodium falciparum* Genome Project [http://www.sanger.ac.uk/Projects/P_falciparum]

19. Stanford Genome Technology Center Malaria Genome Project [http://sequence-www.stanford.edu/group/malaria/index.html]

20. TIGR *Plasmodium falciparum* Genome Database (PFDB) [http://www.tigr.org/tdb/edb2/pfa1/htmls]

21. Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L, Koonin EV, Shallom S, Mason T, Yu K, Fujii C, *et al.*: Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* 1998, 282:1126-1132.

22. Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H: Interpolated Markov models for eukaryotic gene finding. *Genomics* 1999, 59:24-31.

23. Joseph DeRisi lab: web supplement [http://derisilab.ucsf.edu/falciparum]

24. Smith TF, Waterman MS: Identification of common molecular subsequences. *J Mol Biol* 1981, 147:195-197.

25. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, *et al.*: Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 2001, 19:342-347.

26. ArrayOligoSelector [http://arrayoligosel.sourceforge.net]

27. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, *et al.*: Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996, 14:1675-1680.

28. Rouillard J-M, Herbert CJ, Zuker M: OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics* 2002, 18:486-487

29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215:403-410.

30. Jaeger J, Turner DH, Zuker M: Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci USA* 1989, 86:7706-7710.

31. Allawi HT, SantaLucia J Jr: Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. *Biochemistry* 1998, 37:2170-2179.

32. Lyngso RB, Zuker M, Pedersen CN: Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics* 1999, 15:440-445.

33. Peritz AE, Kierzek R, Sugimoto N, Turner DH: Thermodynamic study of internal loops in oligoribonucleotides: symmetric loops are more stable than asymmetric loops. *Biochemistry* 1991, 30:6428-6436.

34. Peyret N, Seneviratne PA, Allawi HT, SantaLucia J Jr: Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches. *Biochemistry* 1999, 38:3468-3477.

35. Sugimoto N, Nakano S, Yoneyama M, Honda K: Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res* 1996, 24:4501-4505.

36. Ziv J, Lempel A: A universal algorithm for sequential data compression. *IEEE Trans Inf Theory* 1977, 23:337-343.

37. Kumar VP, Datta S: Use of variability in the stage-specific transcription levels of *Plasmodium falciparum* in the selection of target genes. *Parasitol Int* 2001, 50:165-173.

38. Certa U, Ghersa P, Dobeli H, Matile H, Kocher HP, Shrivastava IK, Shaw AR, Perrin LH: Aldolase activity of a *Plasmodium falciparum* protein with protective properties. *Science* 1988, 240:1036-1038.

39. Bowman S, Lawson D, Basham D, Brown D, Chillingworth T, Churcher CM, Craig A, Davies RM, Devlin K, Feltwell T, *et al.*: The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* 1999, 400:532-538.

40. Vinkenoog R, Speranca MA, van Breemen O, Ramesar J, Williamson DH, Ross-MacDonald PB, Thomas AW, Janse CJ, del Portillo HA, Waters AP: Malaria parasites contain two identical copies of an elongation factor 1 alpha gene. *Mol Biochem Parasitol* 1998, 94:1-12.

41. Song P, Malhotra P, Tuteja N, Chauhan VS: RNA helicase-related genes of *Plasmodium falciparum* and *Plasmodium cynomolgi*. *Biochem Biophys Res Commun* 1999, 255:312-316.

42. Watanabe J: Cloning and characterization of heat shock protein DnaJ homologues from *Plasmodium falciparum* and comparison with ring infected erythrocyte surface antigen. *Mol Biochem Parasitol* 1997, 88:253-258.

43. Joachimiak MP, Chang C, Rosenthal PJ, Cohen FE: The impact of whole genome sequence data on drug discovery - a malaria case study. *Mol Med* 2001, 7:698-710.

44. Coombs GH, Goldberg DE, Klemba M, Berry C, Kay J, Mottram JC: Aspartic proteases of Plasmodium falciparum and other parasitic protozoa as drug targets. *Trends Parasitol* 2001, 17:532-537.

45. Pinder J, Fowler R, Bannister L, Dluzewski A, Mitchell GH: Motile systems in malaria merozoites: how is the red blood cell invaded? *Parasitol Today* 2000, 16:240-245.

46. McColl DJ, Silva A, Foley M, Kun JF, Favaloro JM, Thompson JK, Marshall VM, Coppel RL, Kemp DJ, Anders RF: Molecular variation in a novel polymorphic antigen associated with *Plasmodium falciparum* merozoites. *Mol Biochem Parasitol* 1994, 68:53-67.

47. de Stricker K, Vuust J, Jepsen S, Oeuvray C, Theisen M: Conservation and heterogeneity of the glutamate-rich protein (GLURP) among field isolates and laboratory lines of *Plasmodium falciparum*. *Mol Biochem Parasitol* 2000, 111:123-130.

48. Patankar S, Munasinghe A, Shoaibi A, Cummings LM, Wirth DF: 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* 2001, 12:3114-3125.

49. van Lin LH, Pace T, Janse CJ, Birago C, Ramesar J, Picci L, Ponzi M, Waters AP: Interspecies conservation of gene order and intron-exon structure in a genomic locus of high gene density and complexity in *Plasmodium*. *Nucleic Acids Res* 2001, 29:2059-2068.

50. Van Dooren GG, Su V, DiOmbrain MC, McFadden GI: Processing of an apicoplast leader sequence in *Plasmodium falciparum*, and the identification of a putative leader cleavage enzyme. *J Biol Chem* 2002, 277:23612-23619.

51.Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 2001, 409:533-538.

52.Horrocks P, Dechering K, Lanzer M: Control of gene expression in *Plasmodium falciparum*. *Mol Biochem Parasitol* 1998, 95:171-181.

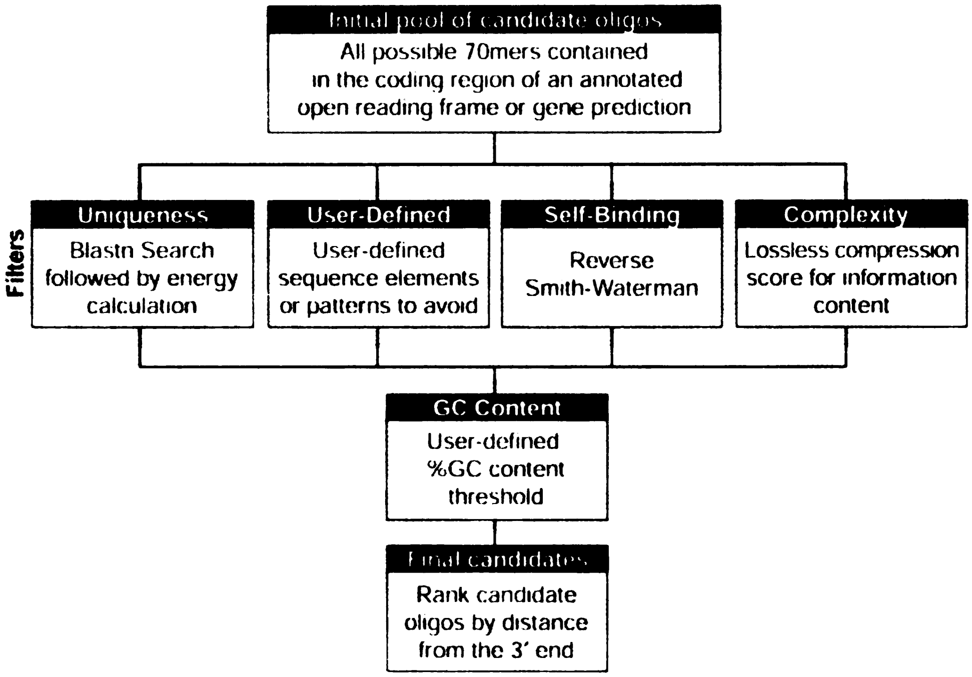
53.Eisen MB, Brown PO: DNA arrays for analysis of gene expression. *Methods Enzymol* 1999, 303:179-205.

54.Trager W, Jensen JB: Human malaria parasites in continuous culture. *Science* 1976, 193:673-675.

55.Microarrays: source for microarray protocols and software [http://derisilab.ucsf.edu/microarray/index.html]

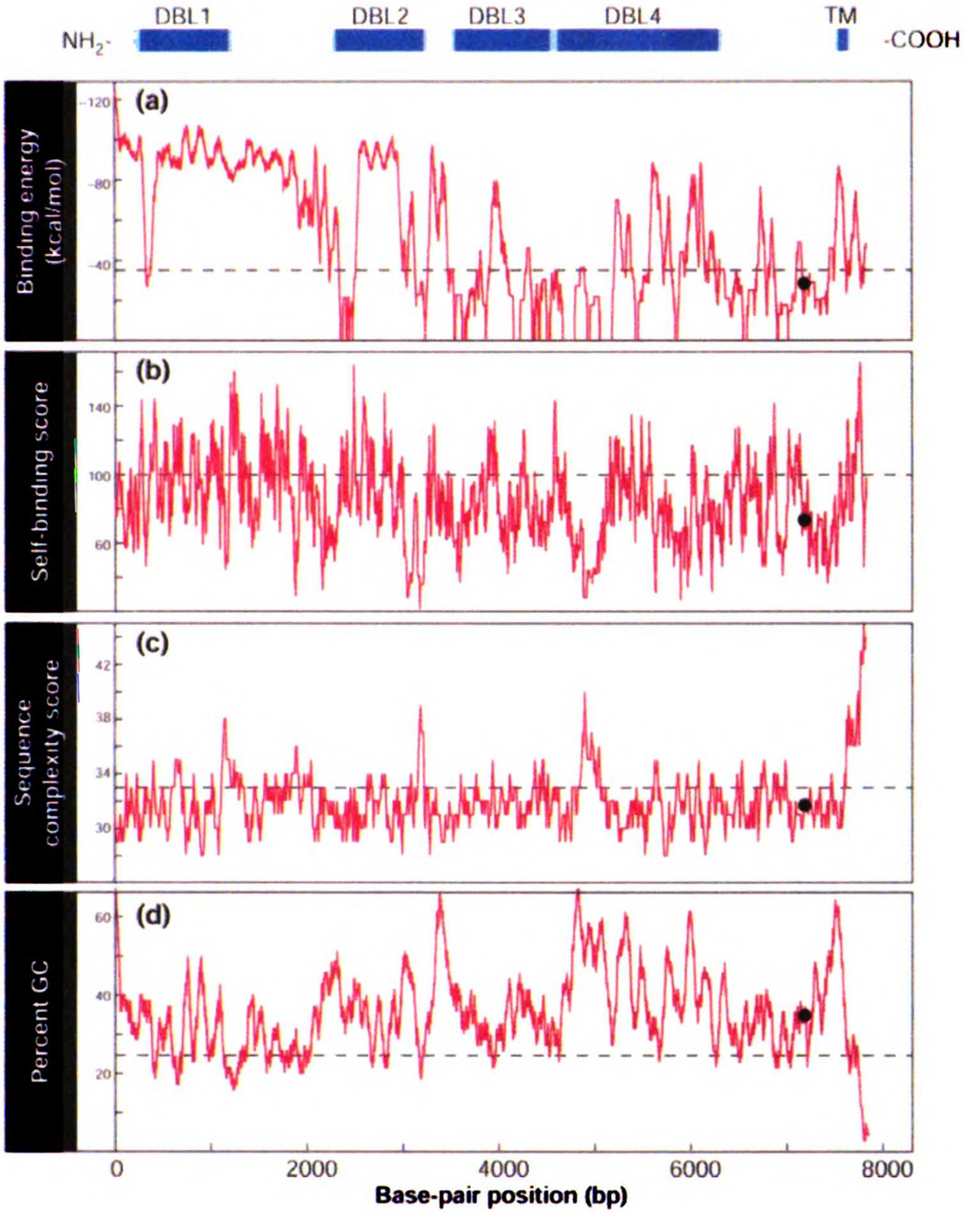
56.Sambrook J, Fritsch EF, Maniatis T: *Molecular Cloning, a Laboratory Manual.* Cold Spring Harbor, NY: Cold Spring Harbor Press; 1989.

Figure 3-1 Schematic of the ArrayOligoSelector oligonucleotide-selection algorithm


Initial pool of candidate oligos
All possible 70mers contained in the coding region of an annotated open reading frame or gene prediction
Filters
Uniqueness
Blastn Search followed by energy calculation
User-Defined
User-defined sequence elements or patterns to avoid
Self-Binding
Reverse Smith-Waterman
Complexity
Lossless compression score for information content
GC Content
User-defined %GC content threshold
Final candidates
Rank candidate oligos by distance from the 3' end

Oligonucleotide selection begins with the collection of all possible 70 mer oligonucleotides from a given ORF. Four filters are executed in parallel: selection for uniqueness within the genome, an optional user-defined pattern filter, avoidance of significant secondary structure (self-binding), and avoidance of low-complexity sequence. The intersection of the set of all oligonucleotides passing these filters are then further selected for a desired base composition and then ranked by proximity to the 3' end of the ORF.

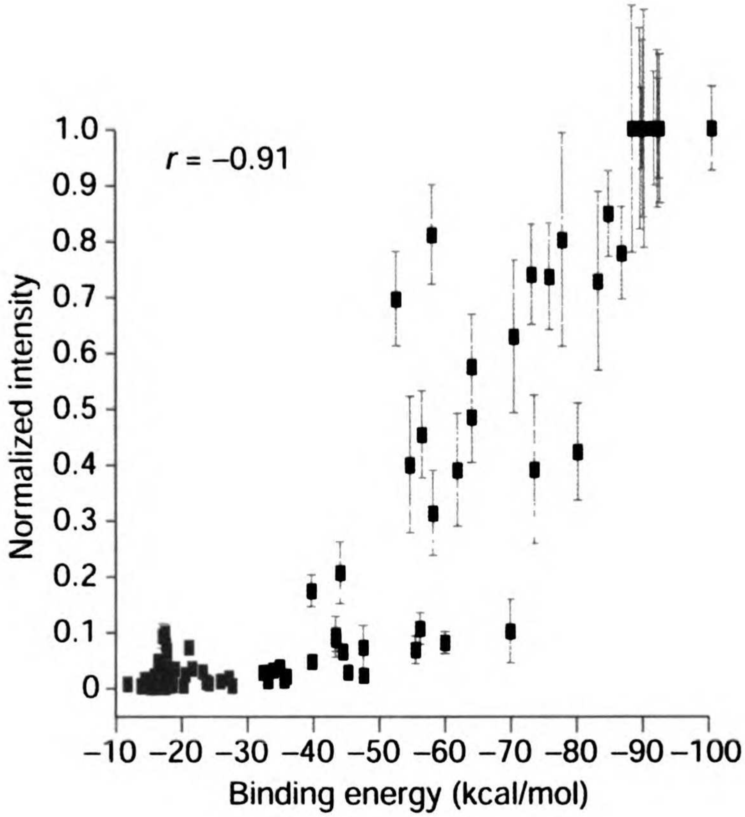
Figure 3-2 Example of sequence parameters measured by ArrayOligoSelector for a putative member of the *var* gene family (PlasmoDB v4.0 annotated gene ID PF08_0140)



A schematic of the target gene is shown above plots for each filter indicating the positions of four Duffy binding-like domains (DBL) and a putative transmembrane domain (TM). For each filter, a 70 bp window was measured for all possible positions

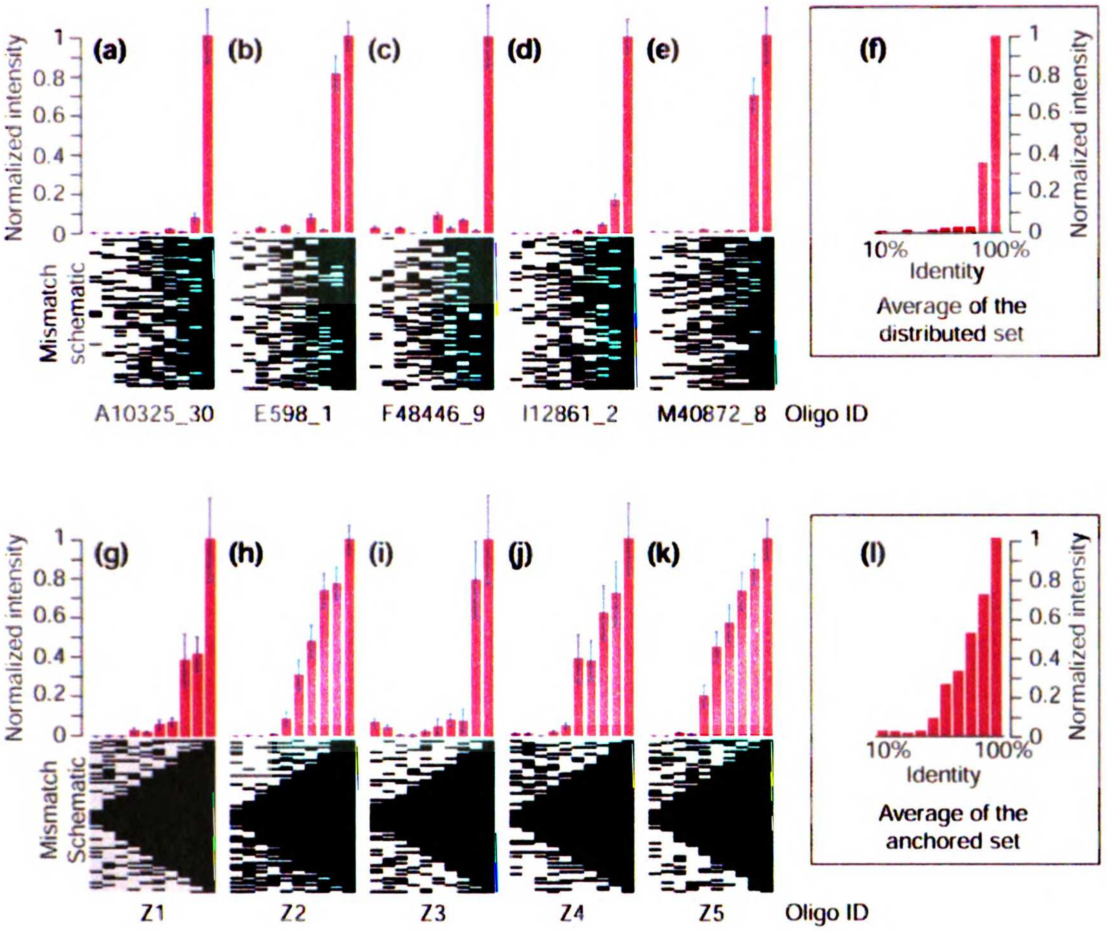
within the gene. The dashed line represents the average value obtained for each filter where 6,000 random 70 bp sequences were chosen from the total collection of *P. falciparum* predicted ORFs. The black circle denotes the position of the final candidate oligonucleotide (oligo ID F44871_2) chosen for the array.

Figure 3-3 Relationship of calculated binding energy to relative hybridization intensity



For each of the oligonucleotides shown in Figure 4, a binding energy (kcal/mol) to the perfect match sequence was calculated using ArrayOligoSelector and plotted against relative hybridization intensity. The Pearson correlation coefficient $|r| = 0.91$ indicates a strong correlation between the calculated binding energy and hybridization performance.

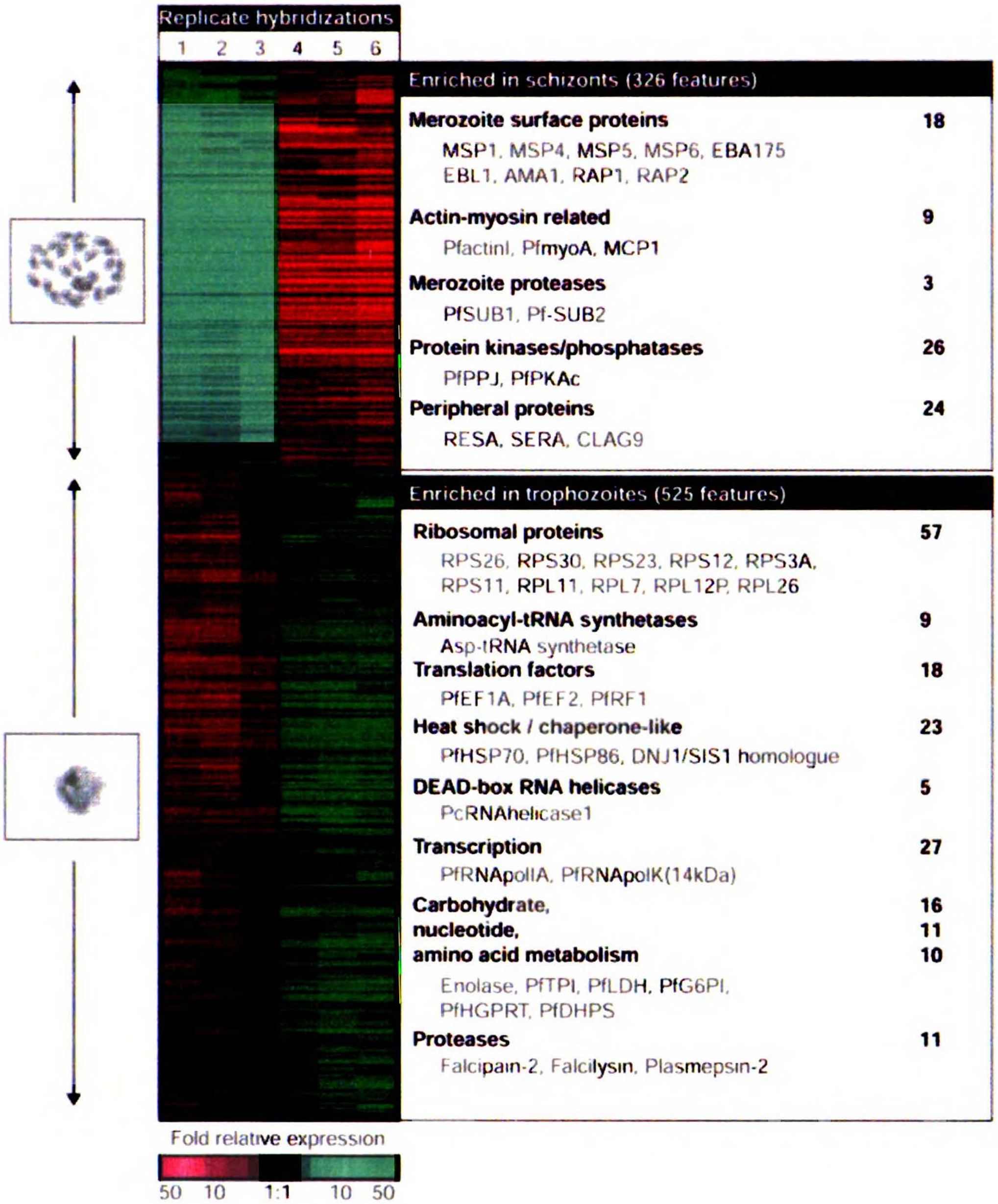
Figure 3-4 Hybridization performance of long oligonucleotides in relation to internal versus terminal mismatches



Ten putative genes were chosen arbitrarily. Each gene is represented by a series of ten oligonucleotides. The first (rightmost) oligonucleotide in each series represents a perfect match to the target sequence. Mismatches were introduced in increasing numbers (10% steps) into each subsequent oligonucleotide in each series. The pattern of matches and mismatches within each oligonucleotide is shown below each graph, where a black bar indicates a perfect match and a gray bar indicates a mismatch. **(a-e)** Mismatch positions were randomized ('distributed set'). **(g-k)** Mismatch positions were chosen such that a

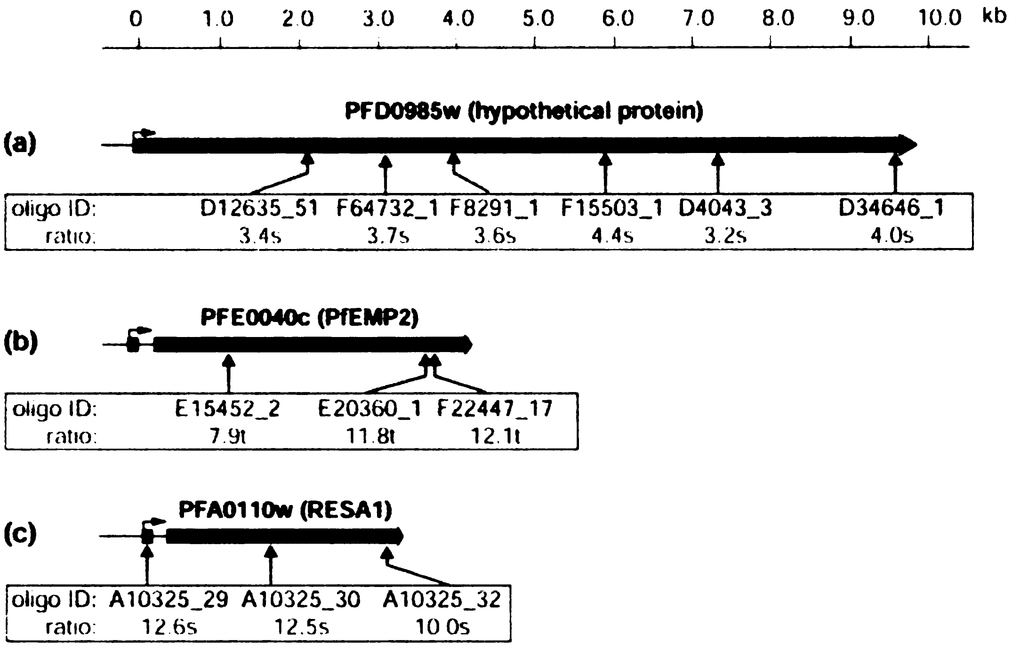
contiguous stretch of perfect-match sequence remained in the middle of the oligonucleotide ('anchored set'). The average hybridization performance, measured as the normalized total intensity for 16 independent experiments, is plotted for each oligonucleotide. **(f)** The average plot for a-e; **(l)** the average plot for g-k.

Figure 3-5 Comparison of trophozoite and schizont stages of *P. falciparum*



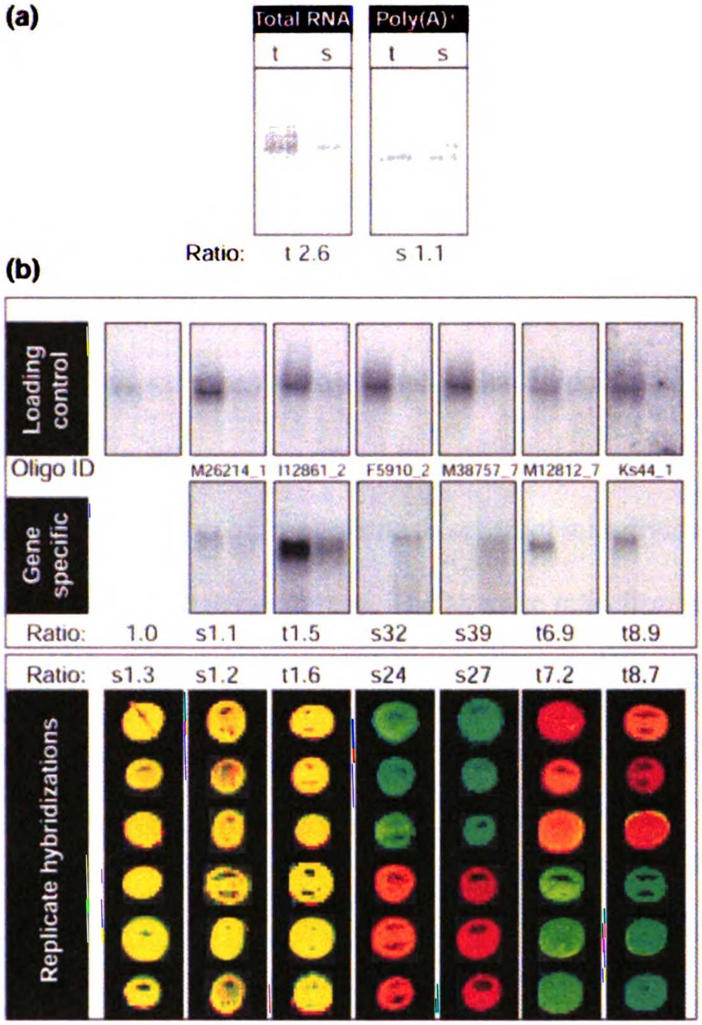
Hierarchical cluster analysis of six replicate microarray hybridizations is shown for the 854 genes that yielded at least a twofold expression difference in at least four experiments. In hybridizations 1-3, the schizont RNA was labeled with Cy3 (green signal) and trophozoite RNA was labeled with Cy5 (red signal) and in analyses 4-6, the fluorophore order was reversed. Examples of major functional gene groups enriched in either studied stage, number of corresponding ORFs and several previously characterized representatives of each group are indicated. All gene expression data are available at PlasmoDB and the DeRisi Lab website [23].

Figure 3-6 Hybridization performance of multiple oligonucleotides representing single ORFs


0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 kb
(a)
PFD0985w (hypothetical protein)
oligo ID: D12635_51 F64732_1 F8291_1 F15503_1 D4043_3 D34646_1
ratio: 3.4s 3.7s 3.6s 4.4s 3.2s 4.0s
(b)
PFE0040c (PfEMP2)
oligo ID: E15452_2 E20360_1 F22447_17
ratio: 7.9t 11.8t 12.1t
(c)
PFA0110w (RESA1)
oligo ID: A10325_29 A10325_30 A10325_32
ratio: 12.6s 12.5s 10.0s

The average expression ratios were calculated for each oligonucleotide overlapping distinct regions of three ORFs: **(a)** PFD0985w encoding a predicted hypothetical protein; **(b)** PFE0040c encoding PfEMP2; and **(c)** PFA0110w encoding RESA1. The arrows indicate the location of oligonucleotide elements within the coding sequence (thick line), and the ratios express fold enrichment in either trophozoite stage versus schizont stage (t) or schizont stage versus trophozoite stage (s).

Figure 3-7 Northern blot validation of the microarray results



(a) Comparison of total RNA and poly(A)$^+$ RNA for northern analysis. 12 µg total RNA and 2 µg poly(A)$^+$ mRNA from both the schizont (s) and trophozoite (t) stages of *P. falciparum* were blotted and probed with a PCR fragment overlapping the genomic sequence corresponding to oligonucleotide ID M11919_1, which represents the 41 kD antigen (p41), fructose-bisphosphate aldolase (PlasmoDB v4.0 annotated gene ID PF14_0425). **(b)** Total RNA northern blots were probed with PCR probes overlapping

the corresponding oligonucleotide elements ('gene specific') for each gene. The selected gene set included a probable calcium-transporting ATPase (PlasmoDB MAL13P1.61), plasmodial heat-shock protein (PlasmoDB PFI0875w), a member of a family of conserved hypothetical proteins (PlasmoDB PFI1445w) and an ORF with sequence similarity to a sodium-and chloride-dependent taurine transporter (PlasmoDB MAL13P1.130), lactate dehydrogenase (PlasmoDB PF13_0141), and a heat-shock protein homolog of Hsp70-3 (PlasmoDB PF11_0351). The northern blot membranes were stripped and reprobed with the fructose bisphosphate aldolase control probe. Northern blot expression levels were measured by a phosphorimager. For each gene-specific probe, the measured ratio of trophozoite to schizont was divided by the ratio measured for aldolase on the same membrane. The average ratio from six replicate hybridizations on the microarray, including fluorophore reversal, is shown in the lower panel. The ratios express fold enrichment in either the trophozoite stage versus schizont stage (t) or schizont stage versus trophozoite stage (s).

Chapter 4 . A Bayesian network driven approach to model the transcriptional response to nitric oxide in *Saccharomyces cerevisiae*

Abstract:

The transcriptional response to exogenously supplied nitric oxide in *Saccharomyces cerevisiae* was modeled using an integrated framework of Bayesian network learning and experimental feedback. A Bayesian network learning algorithm was used to generate network models of transcriptional output, followed by model verification and revision through experimentation. Using this framework, we generated a network model of the yeast transcriptional response to nitric oxide and a panel of other environmental signals. We discovered two environmental triggers, the diauxic shift and glucose repression, that affected the observed transcriptional profile. The computational method predicted the transcriptional control of yeast flavohemoglobin *YHB1* by glucose repression, which was subsequently experimentally verified. To derive Bayesian network models from a combination of gene expression profiles clusters, genetic information and experimental conditions, a software application ExpressionNet was developed and is made freely available.

Introduction:

An extraordinary amount of data has been accumulated measuring genome wide gene expression patterns under a wide range of biological conditions and genetic backgrounds (1, 2). A major challenge in understanding the data is to decipher the relationships between environmental signals, genotypes, cellular phenotypes, protein functions, and the corresponding transcriptional outputs. As the amount of data and the combinations of experimental conditions and genetic backgrounds accumulate rapidly, it becomes increasingly more difficult to form models that seek to explain the observed patterns and relationships.

A systems biology approach combines data mining, computational modeling, and experimental feedback in an iterative cycle of hypothesis generation and testing. Although experimental techniques to generate large-scale microarray data have been well developed, the application of modeling and analysis methods to large data sets continues to mature. Clustering is a commonly used initial approach to analyze gene expression data to identify groups of genes with common or differential expression profiles (3). Many other analysis techniques have also been applied to microarray data such as the signature algorithm to identify both genes and experimental conditions of a co-regulated module (4), algorithms to identify DNA binding motifs using correlation of gene expression (5), mapping gene expression to pathway knowledge (6), and using probabilistic graphic models to construct transcription networks (7). Despite the large amount of data generated through experimental approaches and the increasing application of computational methods to biological problems, the two fields are still by and large

separate. The challenge of applying a systems biology approach lies in the integration of computational modeling and experimental verification.

One computational method employed is a Bayesian belief network. A Bayesian network is a graphic model that encodes probabilistic relationships among variables. The following features make it an attractive framework for modeling the complicated relationships between transcriptional response, experimental conditions and genotypes: (1) decomposition of a large joint distribution over all the variables into independent local relationships; (2) generation of interpretable networks; (3) ability to model causal relationships, hence to gain understanding of a problem domain; (4) ability to handle missing information. A Bayesian network is a graph composed of nodes and edges. The nodes represent variables of interest. The edges represent influence from the parent to child node. The relationships between parent and child nodes are modeled as conditional probability distributions (CPD). Given a set of observed data, a probability score can be assigned to all possible Bayesian network models. We can use Bayesian network learning to search for the most likely network structure given the observed data. The derived graph is a model for the underlying relationships in the data (8, 9).

Bayesian network learning has been applied to infer gene regulatory networks from large scale microarray data by using expression levels of individual genes as network nodes (10-12). In addition, probabilistic graphic models have also been used to integrate heterogeneous data sources to decompose genes into functional modules based on the knowledge of binding motifs, protein interaction and gene expression data (13-15). Although more functionally coherent gene modules were produced in comparison to simple clustering method, the biological conditions and the relationship between those

conditions and the modules were not modeled. More recently, Bayesian networks have been used as the computational framework to predict the genes and the combinatorial constraints that regulate a functional module (16, 17). Many interesting predictions were generated, but only a few were experimentally verified. The separation of computational approach from experimental testing has greatly limited its power to make new biological discoveries.

To overcome this separation, we developed an integrated framework of Bayesian network learning and experimental feedback to model the relationships between the transcriptional response, biological conditions, genotypes, and protein functions. We took advantage of the descriptive power of Bayesian network semantics to formally declare those variables to model microarray data and applied the learning algorithm to elucidate their relationships. Follow-up microarray experiments verified the computational predictions and revealed hidden environmental variables in the original data. Subsequent iterations of modeling and experiments refined the variable declaration and expanded the model with the addition of new experimental data. To circumvent the problems caused by the large data set when inferring gene networks, our approach reduced the number of network nodes by using gene expression cluster profiles instead of individual gene expression levels. This simplified interpretation of the learning result. To elucidate the relationship between transcriptional response and the biological conditions that trigger the response, we extended the network model to account for genotypes, experimental conditions, and protein functions in addition to expression profiles.

We used the framework to investigate the transcriptional response of *S. cerevisiae* to nitric oxide (NO·). Nitric oxide is an important biological agent used by the immune

system to defend against fungal and bacterial infections (18). Studying the yeast response to NO· and the reactive nitrogen intermediates that it generates has important implications to the development of antimicrobial and antifungal treatments. Exposure to chemically generated NO· in yeast triggers both a general stress response as well as a specific NO· detoxification response mediated by the transcription factor Fzf1p (19). The detoxification gene cluster includes the yeast flavohemoglobin *YHB1*, and *SSU1*, a putative sulfite pump in yeast, plus three additional ORFs with unknown functions (19).

To test our approach and to better understand the genome-wide NO· response, we applied the integrative framework to model the transcriptional response to NO· in *S. cerevisiae*. We monitored the genome-wide gene expression of various *FZF1* genotypic strains under multiple experimental conditions including NO· exposure. Using the transcriptome data as input, we generated a network model of the transcriptional response to an extensive panel of environmental perturbations in addition to *FZF1* genotype. The model differentiated the Fzf1p mediated NO· detoxification response from the transcriptional response triggered by other environmental perturbations. We discovered two unappreciated environmental factors that affected the observed expression profiles, and experimentally verified the model prediction that glucose regulates *YHB1* expression.

Results:

Algorithms

Our algorithm iterated through the four steps: data collection and preprocessing, hypothesis generation, model evaluation and experimental feedback (Figure 1).

In the data collection and preprocessing step, gene expression profile clusters were identified from a dataset of microarray experiments and cluster expression levels were converted to discrete values. Each array was annotated with the strain genotype, the experimental conditions, and protein functions if they could be inferred based on the genotype and experimental conditions. The discrete gene expression clusters and the annotated array attributes were combined to form the learning data set.

In the subsequent hypothesis generation step, discrete random variables (network nodes) were defined to model the cluster expression, environmental signals, genotypes and protein functions. Given the learning data set, a probability score could be assigned to a Bayesian network structure. A high scoring network indicates a good fit of the model to the data and the Bayesian network learning process automatically searches for structures with the highest score. The derived model (Bayesian average network) was the average over all the high scoring networks (materials and methods) found by the learning process. Each edge was associated with a confidence score (c), calculated as the percentage of its presence in the high scoring collection (11, 20). The confidence score was a measurement of the data support for an edge.

The derived model was compared to the current biological hypotheses and new predictions were then tested experimentally. In the experimental feedback step, new data was compared to the data underlying the previous model. If the new data conflicted with the previous one, new environmental variables were proposed to seek to explain the discrepancy. The predictions could also be proven incorrect. In either case, we initiated a new iteration of the process to obtain a better random variable definition (i.e. composing a better gene clustering, including new environmental variables) and a better model to 1)

explain the conflict in the data 2) predict the role of the new environmental variables on gene expression 3) eliminate the incorrect predictions.

The initial model

In order to measure the *S. cerevisiae* transcriptional response to NO· and reactive nitrogen intermediates, and to examine the role of the transcription factor Fzf1p, we exposed wild type and *fzf1Δ* strains to chemically generated NO· (experiment E1, materials and methods). To determine whether Fzf1p over-expression could mimic the NO· inducible response, we performed similar experiments with wild type and *GAL1p:FZF1* strains on galactose. We then measured global mRNA levels over time using DNA microarrays (experiment E2, materials and methods). These data were combined with a published dataset from perturbation experiments of yeast treated with common oxidative agents to identify the oxidative or environmental stress response (ESR) (1).

A subset of 130 genes with significant expression changes was selected (materials and methods). We defined five major gene clusters: Fzf1p early and late response clusters (which were up-regulated by NO· in an Fzf1p dependent manner, but differed in their initial response time), the ESR cluster, the oxidative phosphorylation cluster, and the galactose response cluster.

We defined the following ten network nodes and their discrete state values: 1) five gene cluster nodes -"Fzf1p early response", "Fzf1p late response", "ESR", "oxidative phosphorylation" and "galactose response"- for which the change in transcriptional response was modeled as up-regulation, down-regulation and unchanged

expression; 2) three experimental perturbation nodes: "nitric oxide" to model the duration of NO· treatment (0-5, 5-15, 15-45, > 45 min), "galactose" to model the galactose utilization (utilized, not utilized), and "oxidative stress" to model the exposure to common oxidative agents (exposed, not exposed); 3) one genotype node "*FZF1* genotype" (wild type, deletion, over-expression); 4) and one protein function node "Fzf1p activity" to model Fzf1p transcription factor activity (active, inactive).

Given the defined nodes, the learning dataset was composed by combining discrete values of average cluster expression and manual annotation of experimental conditions, *FZF1* genotype and Fzf1p transcription factor activity for each array. The values of Fzf1p activity in the learning dataset were annotated based on *FZF1* genotype and the experimental conditions. For example, if the strain was *fzf1Δ*, the value was assigned to "inactive". Any value that could not be inferred or obtained was set as missing values (empty data entries) in the learning dataset.

The initial derived model (10 edges with $c > 0.9$) is shown in Figure 2a. According to this model, the core NO· specific response (Fzf1p early and late response clusters), unlike other transcriptional responses, was controlled through the activation of the transcription factor Fzf1p. NO· also triggered a general ESR. Those predictions were consistent with the current understanding of the transcriptional response to NO· (19). The remaining model structure was in large part consistent with a manually pre-constructed network structure derived from a biological interpretation of the data (supplemental data).

The edge confidence scores displayed a bimodal distribution with the value 1 or 0 being the most frequent (Figure 2b). This showed a clear separation of edges that were supported or unsupported by the data. This bimodal distribution was vastly different (*P* <

0.001, Kolmogorov-Smirnov normality test) from the normal distribution generated from a collection of networks with randomly assigned structures ($P = 0.35$, Kolmogorov-Smirnov normality test), in which all edges showed a low level of support (0.395±0.105) from data.

As part of the network learning process, CPDs were also computed from the data (supplemental data). This included the node "Fzf1p activity", which had 80% missing values in the learning dataset. The derived CPD of "Fzf1p activity" (the chance of Fzf1p activity in either "active" or "inactive" state given *FZF1* genotype and the duration of NO· treatment) was consistent with all observations and hypotheses of Fzf1p activation triggered by NO· treatment (19) (Figure 2a CPD table).

Two strongly supported edges ($c > 0.9$) in the model were unexpected. One connected from "galactose" to "Fzf1p early response" and the other from "galactose" to "oxidative phosphorylation" (Figure 2a red edges). The CPD of "Fzf1p early response" predicted that the expression of this cluster (containing *YHB1* and *SSU1)* would be up-regulated in response to galactose. Further examination of the microarray data showed *YHB1* was up-regulated by galactose in the absence of Fzf1p over-expression. *FZF1* levels in wild type yeast were not affected by growth in galactose media (Figure 3a wt). Although the new model predicted an Fzf1p-independent up-regulation of the *YHB1* by galactose; it remained a formal possibility that galactose was acting through endogenous Fzf1p to up-regulate *YHB1*.

Experimental feedback and second model

To verify the unexpected *YHB1* induction in response to galactose and the independence of this relationship on Fzf1p, additional expression profiling experiments were performed to monitor the change of mRNA level upon galactose induction in wild type and *fzf1Δ* strains (experiment E3). Indeed, the expression of *YHB1* was increased by 2 to 4 fold upon switching to galactose containing medium (Figure 3b). This agreed with the prediction that galactose affects *YHB1* expression independently of Fzf1p.

In the combined dataset, two galactose induction experiments (experiment E2 vs. E3) were conducted in an experimentally similar way, yet many genes which were up-regulated in one experiment were down-regulated in the other and vice versa (supplemental data). For example, the genes in the Fzf1p response cluster (except *YHB1*) were up-regulated in E2 and down-regulated in E3 (Figure 3a wt vs. 3b wt). In contrast, many galactose utilization genes such as *GAL2*, *GAL3*, *GAL7* and *GAL10* showed consistent up-regulation in all the galactose induction experiments (Figure 3a & 3b). Most of the genes with between-experiment disagreement function to utilize glucose, such as all four subunits of succinate dehydrogenase tetramer *SDH*, acetyl-coA synthetase *ACS1*, and the key gluconeogenic enzymes *FBP1* and *PCK1*. The opposing expression change (E2 vs. E3, wild type) in these glycolysis and gluconeogenesis components were also highly correlated with their transcription profiles during the diauxic shift, the switch from anaerobic growth to aerobic respiration upon depletion of glucose (1, 21). Examining pre-experimental growth conditions, cell densities during the experiment, and the duration of the experiment (12 hr), confirmed that the diauxic shift occurred in the two galactose induction experiments (E2, E3). The diauxic shift also

explained the unexpected connection from "galactose" to "oxidative phosphorylation" node predicted by the initial model (Figure 2a, red edge).

Taken together, the second iteration was initiated to seek a better model to explain the conflict in the data and to predict the effect of the diauxic shift on the gene expression response to NO·. The network nodes were redefined by: 1) the addition of a new environmental factor node "diauxic shift" to model the direction the diauxic shift (entering, exiting, static); 2) redefining the gene cluster nodes as: "ESR", "Fzf1p response", "YHB1", "galactose utilization" and "energy" (materials and methods).

The second model was expanded to take into account the transcriptional response to the diauxic shift (Figure S1a). The new model resolved the conflict in the data and confirmed the connection between the diauxic shift and the energy cluster, and eliminated the relationship between galactose and genes in the glucose utilization pathway.

Glucose derepression and third model

In order to avoid complications due to the diauxic shift in the galactose induction experiments (12 hr), the experiment was repeated using raffinose as the initial sugar source (experiment E4). This allowed a much faster induction and a shorter time course (4 hr). The results showed that the galactose utilization genes such as *GAL7* and *GAL10* were up-regulated; however, *YHB1* induction was not observed (Figure 3c). This result was unexpected since the previous galactose induction experiments showed *YHB1* was induced by 2 to 4 fold (Figure 3a wt, 3b). The difference could not be explained by the diauxic shift or other variables considered thus far.

Growth in glucose rich media represses the transcription of a large number of genes such as enzymes in TCA cycle, the respiratory chain, sporulation genes and genes needed for the utilization of less efficient sugar sources such as galactose (22). To address the possibility that *YHB1* was partially controlled by glucose repression, a node "glucose repression" (repression, derepression) was added to account for this effect in a third model (Figure 2c). The model strongly supported the relationship between glucose derepression and *YHB1* gene expression. The CPD of node "YHB1" predicted *YHB1* gene expression was up-regulated by either glucose derepression or Fzf1p, but not by galactose.

To verify the prediction of glucose derepression on *YHB1*, the protein expression level of GFP tagged Yhb1p was monitored under glucose repression and derepression conditions using flow cytometry. The repression results showed Yhb1p level decreased immediately after the sugar source was changed from either raffinose or galactose to glucose, and continued to decrease up to 2 to 3 fold after 12 hours. This result was confirmed by the reciprocal experiment of glucose derepression by changing the sugar source from glucose to galactose or raffinose, in which Yhb1p level increased by 2 to 4 fold after 12 hours (Figure 4). The ratio and kinetics of the *YHB1* derepression measured by protein level were consistent with the microarray measurements (Figure 3a, 3b). Glucose repression of *YHB1* was not observed in a *TUP1*Δ strain, indicting that the effect of sugar on *YHB1* expression occurred through the canonical glucose repression pathway (data not shown).

Materials and methods:

Microarray experiments

E1: NO· perturbation Log phase (OD_{600} 1.0) strains were treated with NO· released from 1mM DPTA-NONOate (DBY7283, BY4741 *fzf1*Δ) and NO· gas bubbling through the media for 10 seconds (DBY7283). mRNA isolated from treated (DPTA-NONOate exposure for 10, 20, 40, 80, 120 min; 120 min after gas bubbling) or untreated culture was used to generate the Cy5 or Cy3 cDNA probes.

E2: Glucose to galactose I DBY7283 strains with plasmids containing either *GAL1p:LacZ* or *GAL1p:FZF1* were grown to OD_{600} 1.0 in SD-URA, washed by water, then transferred to SGal-URA for continuing growth. mRNA isolated from treated (8, 12 hr after the transfer) or untreated culture was used to generate the Cy5 or Cy3 cDNA probes.

E3: Glucose to galactose II Stationary phase (3 day old saturated glucose culture) DBY7283 and S288c *fzf1Δ* strains were inoculated at OD_{600} 0.5 in SCD, grown for 2 hr, washed by water, then transferred to SCGal for continuing growth. Total RNA isolated from treated (4, 8, 12 hr after the transfer) or untreated cultures was used to generate the Cy5 or Cy3 cDNA probes.

E4: Raffinose to galactose DBY7283 and JZY100 (DBY7283, *FZF1* deleted with KanMX) strains were grown to early log phase in SC raffinose. Galactose was added into the media to a final concentration of 2% for continuing growth. mRNA isolated from sample (0, 30, 60, 120, 240 min after adding galactose) or reference (DBY7283; combined 0 and 240 min) culture was used to generate the Cy5 or Cy3 cDNA probes.

Differentially labeled cDNA probes were hybridized to yeast cDNA microarrays containing PCR probes of all yeast genes (21). Microarray production, RNA isolation,

cDNA synthesis, amino-allyl dye coupling, hybridization and data collection were performed as described (21). Microarray data were normalized using the NOMAD database (ucsf-nomad.sourceforge.net). Spots flagged by GenePix® Pro (Axon Instruments) were excluded from analysis. Spots also excluded from the analysis were both Cy3 and Cy5 signal intensities less than 2 times the background (E1, E2) and with feature intensity less than the background (E3, E4). E4 dataset was transformed (i.e. normalized) by its 0 min data point. Complete microarray data are available at http://derisilab.ucsf.edu/network.

Data source and preprocessing for network learning

Initial model The microarray data included those from experiments E1 or E2 (19 arrays) and the published dataset of yeast treated with H_2O_2 or menadione over 0 to 160 min (21 arrays) (1). A subset of 130 genes with greater than 2 fold change in 3 or more data points in the E1 and E2 experiments were selected. The genes were clustered using data from experiments E1 and E2 (3). 5 major gene clusters were identified using correlation cutoff 0.75 with subsequent manual adjustment: Fzf1p early response, Fzf1p late response, ESR, oxidative phosphorylation, and galactose response clusters. The manual adjustment consisted of combining galactose up-regulation and down-regulation clusters to galactose response cluster and splitting Fzf1p response cluster into the Fzf1p early and late response clusters based on their initial response times (5-15 min vs. 15-45 min).

Second model The microarray data included those in the initial modeling plus array data generated from experiment E3 and a published dataset monitoring the transcriptional response of the diauxic shift (21). The 130 genes selected in the initial modeling were clustered using all the above microarray data (3). Five major gene clusters were identified

using cutoff 0.6 with subsequent manual adjustments: ESR, Fzf1p response, YHB1, galactose utilization and energy clusters. The energy cluster included genes in the previously designated oxidative phosphorylation cluster and genes in the glucose utilization pathway that were previously in the galactose response cluster. The manual adjustments consisted of separating *YHB1* from the Fzf1p response cluster and forming a YHB1 cluster contained only *YHB1*. After obtaining the new data, we realized the crucial separation within the Fzf1p response cluster was that between *YHB1* and the rest of the cluster. Therefore, the Fzf1p response clusters were not separated into early and late response clusters as in the initial model.

Third model The microarray data included those in the second model as well as those generated from experiment E4. Gene clusters were defined using all the above microarray data with the same procedure as those used in the second model.

The learning dataset was composed of 1) average gene expression change of each gene cluster converted into discrete values using a 2-fold threshold and 2) manual annotation of the remaining values (experimental perturbations, *FZF1* genotype and Fzf1p activity) based on the experimental conditions and strain genotypes (missing values were allowed). The complete learning datasets and gene membership of each cluster are available at http://derisilab.ucsf.edu/network.

Bayesian network learning and software implementation

A software application, ExpressionNet, was developed to perform Bayesian network learning. We used a Bayesian scoring function to assign a probability score for a network model. The clique-tree technique as well as the variable-elimination technique were implemented for efficient inference and learning (9, 23). The learning process

started with random edge combinations, gradually improving the network topology using a greedy search strategy until the score reached a local maximum. The greedy search was iterated to generate a collection of high scoring networks. High scoring networks were subjected to small topology changes by single edge addition, deletion or reversion to expand the collection. Learning was repeated using two different prior probability distributions of the network parameters (priors), both set as a Dirichlet distribution: $Dir(1,1, \ldots, 1)$ and $Dir(P_0 \cdot \alpha, P_0 \cdot \alpha, \ldots, P_0 \cdot \alpha)$, where P_0 is a uniform distribution over the probability space of each CPD and $\alpha=5$. Networks scoring within a percentile cutoff (15% for the initial and second models, 25% for the third model) using both priors were used to construct average Bayesian network models. We defined all environmental and genotype nodes as root nodes and all gene cluster nodes as leaf nodes. Missing values was handled using a Structural Expectation-Maximization algorithm (24). ExpressionNet is available at http://expressionnet.sourceforge.net/. The derived network models and probability parameters are available at http://derisilab.ucsf.edu/network.

Flow Cytometry

The *YHB1-GFP* strain is a C-terminus fusion of *GFP* obtained from a genome-wide tagged library (25). The culture was grown to early log phase in synthetic media with 2% glucose, raffinose or galactose, washed with PBS, then transferred to synthetic medium with 2% glucose (from raffinose or galactose), or raffinose or galactose (from glucose). The cell fluorescence intensities were measured on a Becton Dickison LSR II flow cytometer at 0, 2, 4.5, 6, 8.25, or 12 hr after the sugar was changed. For each time point, a minimum of 100,000 cells were measured to derive the mean GFP intensity.

Discussion

We developed a framework to formally couple Bayesian network learning and experimental feedback to successfully model a specific biological response in yeast. Computational modeling formalized biological hypotheses in a probabilistic language and generated models to elucidate the relationships between environmental perturbations and the transcriptional output. Experimental feedback verified new computational predictions and revealed additional environmental factors, which were essential for the refinement and expansion of the model. We were able to use this integrative approach to achieve two goals. First, we discovered a relationship that had been difficult to recognize by using either computational or experimental approach. Secondly, our approach dissected out specific versus nonspecific responses to NO· and reactive nitrogen intermediates. The core structure of the Fzf1p-dependent NO· specific response sub-network (nitric oxide, *FZF1* genotype Fzf1p activity, and Fzf1p response clusters) was predicted and maintained throughout the three models. The transcriptional responses to other environmental factors were gradually elucidated by additional iterations of the process.

Previous studies have suggested that *YHB1* is important for the survival of yeast under oxidative and nitrosative stress (26, 27). Our results showed *YHB1* was transcriptionally regulated by both NO· exposure and glucose repression. Taken together, these data indicated that *YHB1* is regulated by many environmental signals, highlighting the combinatorial control of this gene. While glucose derepression caused a 2 to 3-fold increase in Yhb1p protein level, studies have shown a 10-fold increase by NO· treatment, suggesting a more prominent role of Yhb1p in NO· detoxification (19).

A common practice in statistical learning is to select one single model that best fits the data. But in many situations, other models also score very well although not necessarily the best. Using the best single model to derive a biological conclusion is potentially risky due to data over-fitting. To circumvent this problem, the average of all the high scoring networks was found by the searching procedure (20). An added benefit of this approach is that it yields a confidence score associated with each edge connection (11). The confidence score is especially useful for filtering low-confidence connections from complex networks, thus simplifying what might otherwise be a confusing network. In addition to network averaging, the high scoring networks were the convergence of learning using two different priors. This strategy overcame network structure bias caused by using a single prior.

The edges in a Bayesian network represent the influence from parent to child nodes. These are statistically favorable solely based on the data and priors. After model averaging, meaningful biological connections may be inferred from the high confidence edges. Since Bayesian network edges represent statistical instead of causal relationships, it is possible a derived edge does not represent a direct biological connection. For example, two gene clusters sharing high mutual information would likely be connected. One method to eliminate such connections is to merge those highly correlated clusters into a single node. Additionally, structural constraints may be used to define gene expression nodes as leaf nodes and the environmental variable nodes as root nodes.

Gene clusters were defined through an automatic hierarchical clustering algorithm with manual interventions. Those manual interventions were comprised of the selection of correlation cutoffs and sub-cluster to node assignments. Although it is not purely

automatic, this step is critical for ensuring the quality of gene cluster node definitions, and therefore critical for producing high quality network models. In our future work, we will generate gene clusters automatically and integrate clustering with Bayesian network learning to take advantage of the network structure to optimize both clustering and network models (28).

This computational framework is not limited to microarray gene expression data. It can be extended to incorporate data such as protein expression, genetic interactions, and sequence motifs. As the datasets grows larger and more complex, tightly coupled computational modeling and experimental feedback provides an efficient approach to study a biological system.

Reference

1. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D. & Brown, P. O. (2000) *Mol Biol Cell* **11,** 4241-57.

2. Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. & Friend, S. H. (2000) *Cell* **102,** 109-26.

3. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc Natl Acad Sci U S A* **95,** 14863-8.

4. Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. & Barkai, N. (2002) *Nat Genet* **31,** 370-7.

5. Bussemaker, H. J., Li, H. & Siggia, E. D. (2001) *Nat Genet* **27,** 167-71.

6. Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C. & Conklin, B. R. (2002) *Nat Genet* **31,** 19-20.

7. Friedman, N. (2004) *Science* **303,** 799-805.

8. Pearl, J. (1988) *Probabalistic reasoning in intelligent systems: networks of plausible inference* (Morgan Kaufmann Publishers, San Mateo, Calif.).

9. Jordan, M. I. (1999) *Learning in graphical models* (MIT Press, Cambridge, Mass.).

10. Pe'er, D., Regev, A., Elidan, G. & Friedman, N. (2001) *Bioinformatics* **17,** S215-24.

11. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000) *J Comput Biol* **7,** 601-20.

12. Hartemink, A. J., Gifford, D. K., Jaakkola, T. S. & Young, R. A. (2001) *Pac Symp Biocomput,* 422-33.

13. Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. & Botstein, D. (2003) *Proc Natl Acad Sci U S A* **100,** 8348-53.

14. Segal, E. & Koller, D. (2003) *Bioinformatics* **1,** 1-9.

15. Segal, E., Yelensky, R. & Koller, D. (2003) *Bioinformatics* **19 Suppl 1,** I273-I282.

16. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. & Friedman, N. (2003) *Nat Genet* **34,** 166-76.

17. Beer, M. A. & Tavazoie, S. (2004) *Cell* **117,** 185-98.

18. MacMicking, J., Xie, Q. W. & Nathan, C. (1997) *Annu Rev Immunol* **15,** 323-50.

19. Sarver, A. & DeRisi, J. L. (2005) *Submitted.*

20. Hoeting, J., Madigan, D., Raftery, A. & Volinsky, T. (1999) *Statistical Science* **14,** 382–417.

21. DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997) *Science* **278,** 680-6.

22. Broach, J. R., Pringle, J. R. & Jones, E. W. (1991) *The Molecular and cellular biology of the yeast Saccharomyces* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.).

23. Heckerman, D., Geiger, C. & Chickering, D. (1995) *Machine Learning* **20,** 197-243.

24. Friedman, N. (1998) in *Uncertainty in Artificial Intelligence. Proceedings of the Fourteenth Conference*, ed. S., C. G. M. (Morgan Kaufmann Publishers, Madison, WI, USA), pp. 129-38.

25. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S. & O'Shea, E. K. (2003) *Nature* **425,** 686-91.

26. Liu, L., Zeng, M., Hausladen, A., Heitman, J. & Stamler, J. S. (2000) *Proc Natl Acad Sci U S A* **97,** 4672-6.

27. Zhao, X. J., Raitt, D., P, V. B., Clewell, A. S., Kwast, K. E. & Poyton, R. O. (1996) *J Biol Chem* **271,** 25131-8.

28. Segal, E., Battle, A. & Koller, D. (2003) *Pac Symp Biocomput*, 89-100.

Figure 4-1. Illustration of the iterative network learning and experimental feedback algorithm

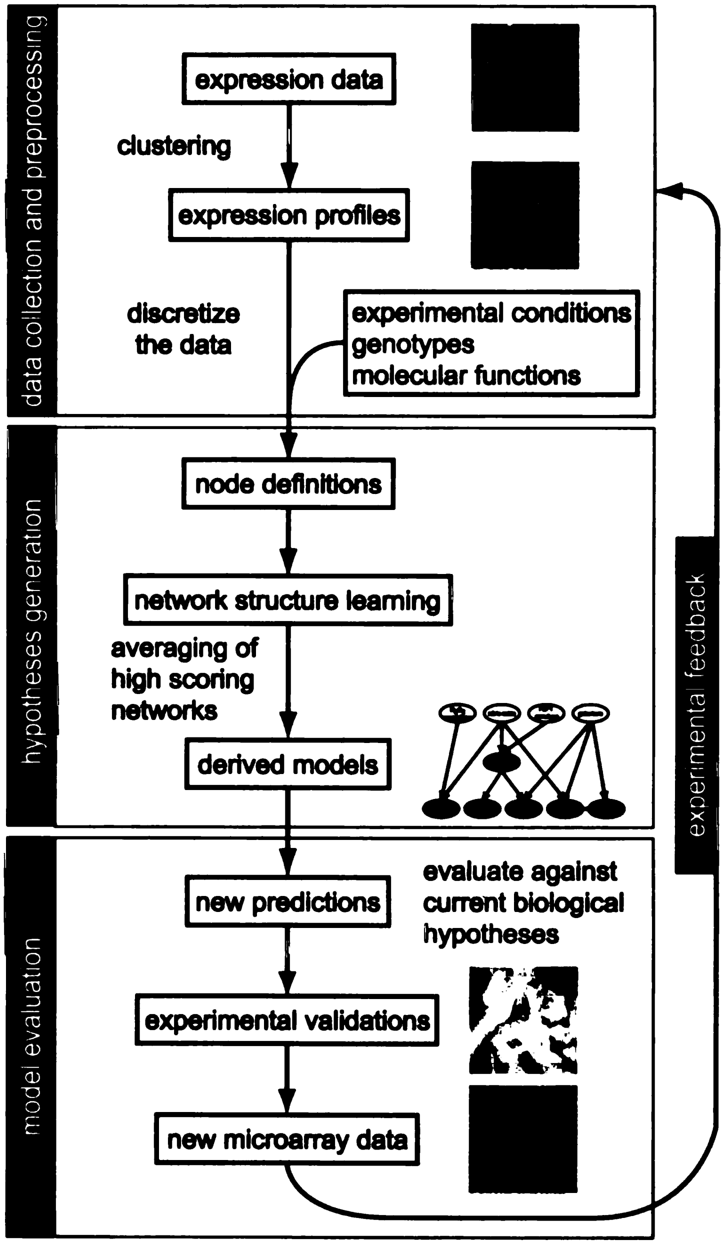
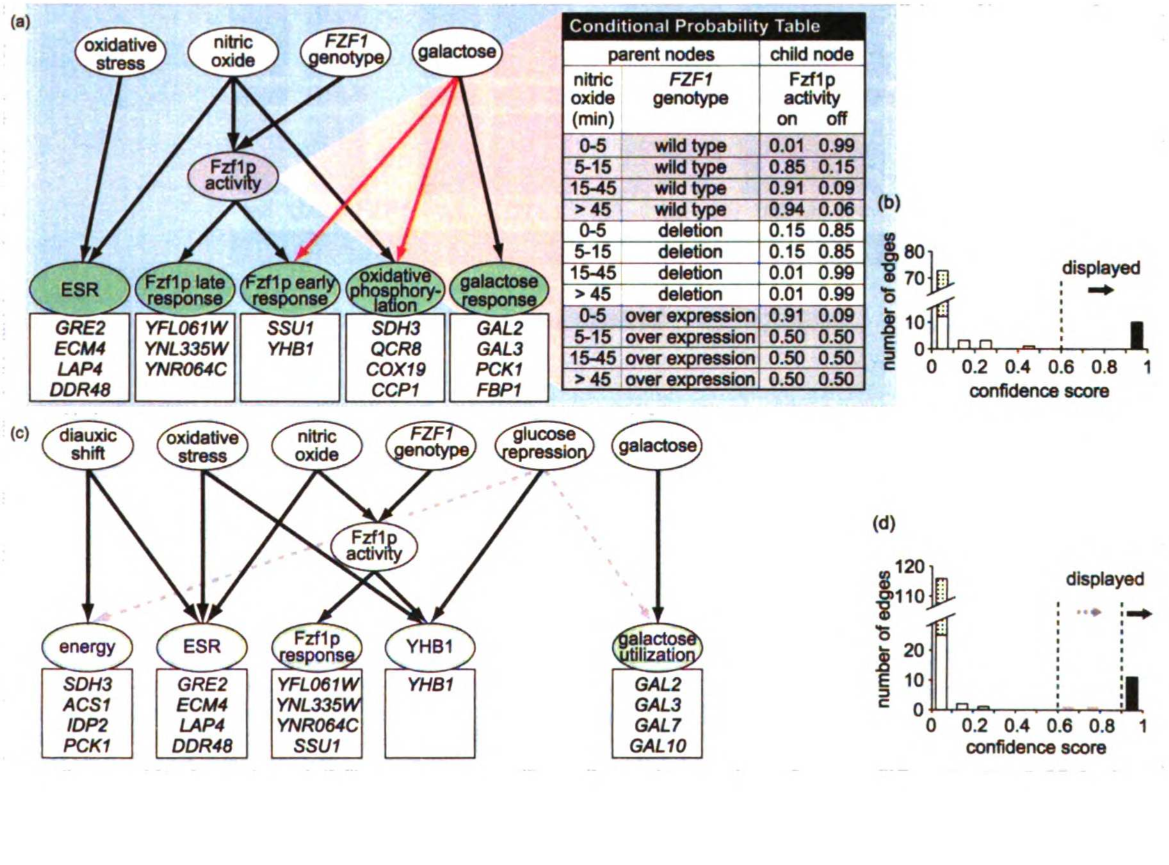


Figure 4-2 The Bayesian average network representation of the models

(a, b) initial model (c, d) third model (a, c) Graphic representation. The green nodes represent gene expression clusters. Representative genes of each cluster are shown in the box below each node. ESR: environmental stress response cluster. Energy: glucose metabolism cluster. Oxidative stress: the application of H_2O_2 or menadione. Nitric oxide: the duration of NO· exposure. Galactose: galactose utilization. Diauxic shift: direction of the diauxic shift. Nodes with missing values are colored in gray. The CPD table shows the conditional probability distribution of Fzf1p activity. The red edges represent novel predictions from the network model. (b, d) Edge confidence score histogram. The dot-filled columns represent edges excluded from the model based on structural constraints.



Conditional Probability Table

parent nodes		child node	
nitric oxide (min)	*FZF1* genotype	Fzf1p activity on	off
0-5	wild type	0.01	0.99
5-15	wild type	0.85	0.15
15-45	wild type	0.91	0.09
> 45	wild type	0.94	0.06
0-5	deletion	0.15	0.85
5-15	deletion	0.15	0.85
15-45	deletion	0.01	0.99
> 45	deletion	0.01	0.99
0-5	over expression	0.91	0.09
5-15	over expression	0.50	0.50
15-45	over expression	0.50	0.50
> 45	over expression	0.50	0.50

Figure 4-3 The change of gene expression of Fzf1p response cluster, *FZF1* and galactose utilization genes in response to galactose

(a) Wild type and gal promoter driven *FZF1* over-expression strains in response to change from glucose to galactose (experiment E2). (b) Wild type and *fzf1*Δ in response to change from glucose to galactose (E3). (c) Wild type and *fzf1*Δ in response to change from raffinose to galactose (E4). Color unit is fold change of gene expression. Gene expressions are too low to detect are colored in blue.

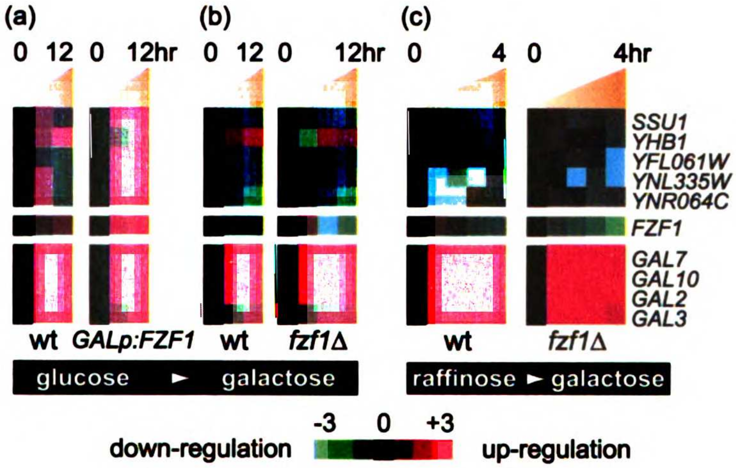


Figure 4-4 Glucose repression and derepression of Yhb1p-GFP measured by flow cytometry

To calculate a mean GFP intensity, a minimum of 100,000 cells were measured for each time point.

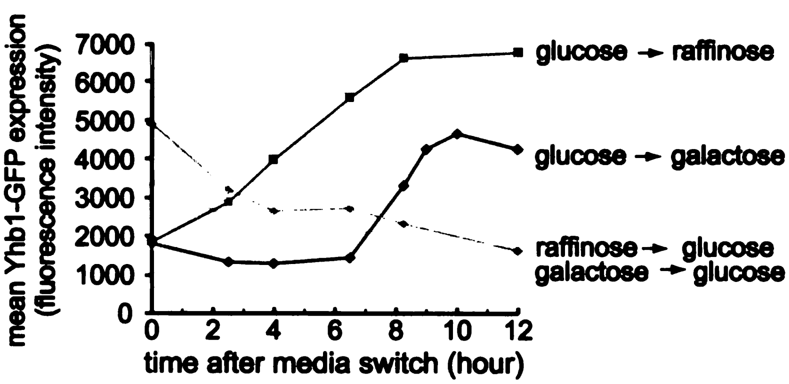
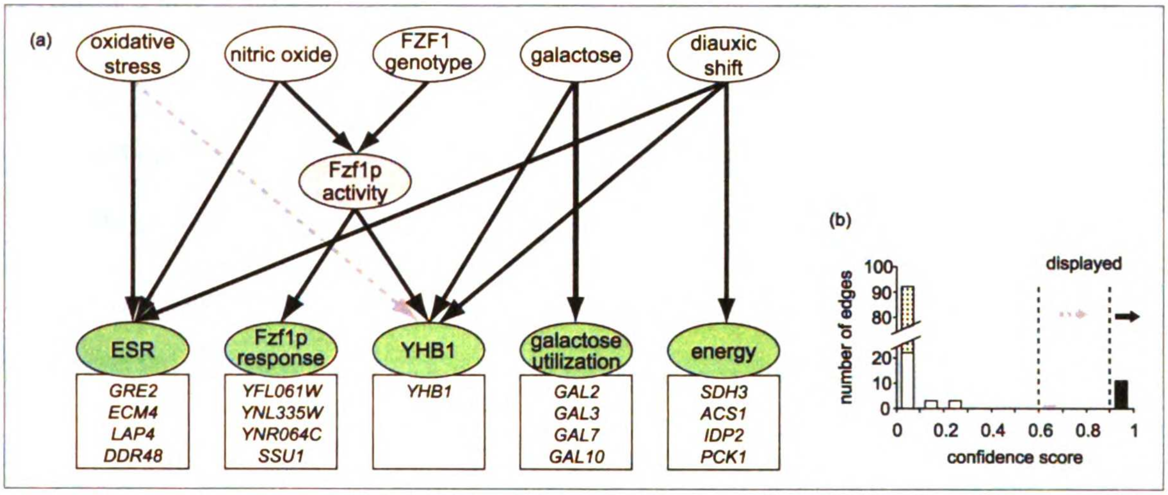


Figure 4-S1 The Bayesian average network representation of the second model

(a) Graphic representation. The green nodes represent gene expression clusters. Representative genes of each cluster are shown in the box below each node. ESR: environmental stress response cluster. Energy: glucose metabolism cluster. Oxidative stress: the application of H_2O_2 or menadione. Nitric oxide: the duration of NO· treatment. Galactose: galactose utilization. Diauxic shift: direction of the diauxic shift. The node with missing values is colored in gray. (b) Edge confidence score histogram. The dot-filled column represents edges excluded from the model based on structural constraints.



Chapter 5 . Virtual Drug Development for New Antimalarials

INTRODUCTION

Malaria is one of the most deadly infectious diseases in the world, causing approximately 350- 500 million clinical episodes and as many as three million deaths annually. As of 2004, 107 countries reported areas at risk of malaria transmission, mostly developing countries. In part of Africa, the estimated infection rate is on average once every two years per person. WHO has estimated that 80% of deaths attributed to malaria occur in African children under the age of five [1-3].

Malaria is transmitted by mosquitoes but its causative agent is a type of intracellular protozoan parasite from the genus *Plasmodium*. There are many species of *Plasmodium*, but only four (*Plasmodium ovale, Plasmodium vivax, Plasmodium malariae, Plasmodium falciparum*) cause malaria in humans. Of the four species, *P. falciparum* has the most severe morbidity and mortality rates [4].

In developed countries, malaria has been virtually eradicated through mosquito control, yet the disease still plagues most of the developing world in the tropical and subtropical regions [1,5]. Currently, once a person has been infected with the parasite, effective treatments are antimalarial drugs that were developed during or after World War II. Most of the drugs to treat and/or prevent malaria belong in four categories: quinine and its derivatives, antifolate combination drugs, artemisinin compounds, and tetracycline and its derivative antibiotics. Chloroquine is a 4-aminoquinoline derivative of quinine, developed in the 1930s and 40s. Since it is very inexpensive to synthesize, chloroquine

has been historically the most widely used antimalarial. Other quinine family drugs include quinidine, amodiaquine, primaquine and mefloquine. Even though it has been the most widely used, the mechanism by which chloroquine kills the parasites is still unclear, possibly by interfering with the parasitic heme polymerization process [6]. Antifolate combination drugs are various combinations of dihydrofolate reductase inhibitors such as pyrimethamine and sulfadoxine [7]. The sweet wormwood (also known as Qinghao in Chinese) has been used by the Chinese for thousands of years to treat malaria, though its active compound, artemisinin, was discovered only in the 1970s, working possibly through generating peroxide radicals or forming Fe(IV)=O species as the toxic agent [8,9]. Tetracycline and its derivative antibiotics such as doxycycline are potent antimalarials and are used for both treatment and disease prevention [7].

Unfortunately, malaria parasites have developed resistance to most of the commonly used drugs such as chloroquine and the resistance has been spreading quickly throughout epidemic regions [1,7,10]. This has been one of the key factors for the failure of programs by the World Health Organization (WHO) to eradicate and/or control the spread of disease [3,11]. Chloroquine resistance initially developed in parts of Southeast Asia in the 1960s and quickly spread to all major malaria transmitting regions except Africa. By the 1980s, the resistant strains finally reached the African continent. By 2005, only a few areas in South America still have chloroquine-sensitive parasites [1,12]. Resistance to antifloate compounds (sulfadoxine and pyrimethamine) is found frequently in Southeast Asia and South America. Even more alarming is the occurrence of multi-drug resistant *Plasmodium* strains in Southeast Asia, which if it spreads widely, can be devastating to the world [7,10,13-16].

With such a limited choice of available drugs, there is an urgent need to develop new antimalarials. This includes both identifying new drug targets, developing new antimalarial compounds, and testing existing drugs for reuse as antimalarials. Such a repurposing example is triclosan, a common ingredient in toothpaste, soap and shampoo for its antimicrobial properties. Recent studies have shown triclosan exhibiting potent inhibition to malaria growth, working through the same inhibitory mechanism as an inhibitor for type II fatty acid biosynthesis [17-19].

The complete genomes of *P. falciparum* and *P. yoelli yoelli* became available in 2002 and sequencing of other *Plasmodium* species is underway [4,20,21], which provides us a complete parasitic gene list to identify potential new drug targets. At the mean time, the number of organisms whose genomes have been completely sequenced has also increased dramatically [22]. The wealth of genomic information provides us a genomic opportunity to identify drug candidates that are most likely to be effective: essential for the parasites and lacking a significant human homolog [22]. This property can be seen in targets of many commonly used antibiotic targets, such as penicillin targeting peptidoglycan transpeptidase, an essential enzyme for synthesizing bacterial cell walls [23]. It is still very difficult to experimentally validate the essentiality of each *Plasmodium* gene; conservation through evolution can be used as a proxy to essentiality. Using these principles, comparative analysis of the *Plasmodium* genome with many other genomes including mammalian, plant, fungal, microbial and archaeal genomes can identify potential drug targets by finding malaria genes that are conserved through evolution but not represented or represented in a very distinct form in the human genome. A likely place to find such genes is the parasite apicoplast, a relic chloroplast of

prokaryotic origin, which harbors enzymes and metabolic pathways that are shared by plants and prokaryotes [18,24].

In addition to genome sequence information, there are many resources and computational tools that can further lead us from drug target genes to identifying small molecule inhibitors. For example, comparative protein structure modeling could be used to predict 3D structures of the drug target proteins [25]; *P. falciparum* metabolic pathway databases to extract substrate information of drug target enzymes [26]; molecular docking to identify small molecules targeting specific proteins [27]; molecular profiling to eliminate molecules without "drug-like" properties [28]; and structure-based similarity searches for virtual screens of chemical databases [29].

In this manuscript we present an informatics approach that started with genome sequence information and led to identifying potential antimalarial compounds. We used this approach to systematically mine the entire *P. falciparum* proteome for potential drug targets, and further integrate comparative protein structure modeling, enzymatic annotations, molecular profiling and virtual screening of chemical databases to identify small molecular inhibitors. We identified 152 potential antimalarial drug targets, 77 ligands that potentially bind to the drug targets, and 1893 commercially available small molecules with potential antimalarial activities, which will be tested experimentally.

MATERIALS AND METHODS

Proteome collection

795,685 annotated protein sequences from 203 completely sequenced genomes (including *P. falciparum*) were used to generate the phylogenomic profiles. The genomes are listed in supplemental Table S1.

Phylogenomic profile matrixes

By sequence homology

Two phylogenomic profiles were constructed using the sequence homology scores as comparison metrics (E values or bits scores). We blasted *P. falciparum* protein sequences against 795,685 protein sequences from 202 organisms (BLASTP 2.2.3: E-value threshold = 1.0, low complexity filter on, default values were used for the rest of program parameters) [30]. The best E-value ($\log_{10}$E-value) or bit score for each *P. falciparum* protein within each genome was extracted to construct the phylogenomic profiles matrix. The matrix was arranged in a way that each row represented a *P. falciparum* protein and each column represented a genome. The phylogenomic profile matrixes were clustered on both the *P. falciparum* proteome and genome directions.

By sequence homology of orthologs

Two phylogenomic profiles were constructed by sequence homology of orthologs (either E values or bits scores). A bi-directional blast of *P. falciparum* genome to other genomes in the proteome collection was performed using the same BLASTP parameters as described before. Orthologs were identified as the best reciprocal BLASTP hits (by either E-value or bit scores) between two genomes. When orthologs were identified, $\log_{10}$E-value or bit scores generated by blasting *P. falciparum* proteome to other

proteomes were extracted to construct the phylogenomic profiles matrix with the same method as described before.

By similarity to protein domain and family signatures

Protein signature profiles from InterPro and its following member databases -- Pfam, ProDom, SUPERFAMILY, SMART, PIRSF, and PRINTS -- were obtained from InterPro release 8.0 [31]. TIGRFAMs protein family profiles were obtained from TIGRFAMs release 4.0 [32]. Profile assignments and associated similarity scores to Swiss-Prot and TrEMBL protein sequences [33] were extracted from InterPro 8.0. TIGRFAMs profile assignments of *P. falciparum* protein sequences and similarity scores were calculated using the InterProScan program [34]. Similarity scores were $\log_{10}$E-values for all profile assignments except for the matches to InterPro entries, which were binary values (-1 and 0) with -1 representing a match in order to conform to the sign of $\log_{10}$E-value.

For a given profile, the best score ($\log_{10}$E-value or binary values) in a genome was used to construct the phylogenomic profile matrixes. A single matrix was constructed for profiles defined in a single database, in which a row represented a profile and a column represented a genome.

Drug target scoring function

For each *P. falciparum* protein, a score was calculated for each phylogenetic comparison method (corresponding to each phylogenomic matrix). The score had two components. The first component ($S_{io} \bullet C_{io}$, see below) measured the conservation in

non-mammalian genomes, penalized by a second component ($S_{im} \bullet C_{im}$, see below), which measured the existence of homologs in mammalian genomes. The final drug target score (*S*) for a *P. falciparum* protein was defined as the weighted sum of scores derived from individual matrix.

$$S = \sum_i w_i(S_{io} \bullet C_{io} - S_{im} \bullet C_{im})$$

i is the index of each phylogenetic comparison method. w_i is the weight for each method (equal weight was used $w_i = 1$).

For sequence homology based methods (4 total), profiles of a *P. falciparum* protein were extracted from matrix *i*. S_{io} is the average value, in profile *i*, of non-mammalian genomes whose values exceed the cutoff (bit score >100, $\log_{10}$E-value <-10). C_{io} is the fraction of the non-mammalian genomes with values exceeding the cutoff, in profile *i*. S_{im} and C_{im} are defined in the same way for mammalian genomes (cutoffs: bit score <0, $\log_{10}$E-value >0).

For protein domain or family signatures based method (8 total), signatures were first mapped to *P. falciparum* genes if the signature could be detected in the protein sequence. The signature profiles were extracted from the phylogenomic matrix *i* and then transferred to *P. falciparum* genes when such mappings were available. For each *P. falciparum* gene and each comparison method *i*, S_{io} is the expectation value of the signature *i* detected in the *P falciparum* gene. C_{io} is the fraction of the non-mammalian genomes in profile *i* with values exceeding cutoff ($\log_{10}$E-value <–0.1). Sum of $S_{io}.C_{io}$ was used in place of $S_{io}.C_{io}$ if multiple signatures (all come from method *i)* mapped to a single gene. S_{im} and C_{im} were defined in the same way for mammalian genomes (cutoff: $\log_{10}$E-value >0). For InterPro entries, binary values (-1, 0) were used in place of E-

values for phylogenomic profile matrix construction as well as S_{io} , C_{io}, S_{im} and C_{im} calculations (cutoffs: <-0.5 for S_{io} , C_{io}, and >-0.5 for S_{im} , C_{im},).

Ligand identification based on MODBASE/LIGBASE predictions

MODBASE (generated by ModPipe program on 2003-10-1) and LIGBASE predictions for *P. falciparum* proteome were supplied by the Sali lab, UCSF. The reliability of a predicted model generated by comparative protein structure modeling can be evaluated by a reliability score [35]. A model is predicted to be reliable in MODBASE (95% chance having at least 85% of its C alpha atoms superposed within 3.5Å of their correct positions) when the score is higher than 0.7. Using 0.7 as the cutoff, 2019 *P. falciparum* proteins were identified in MODBASE to have 3276 reliably predicted structure models [36,37].

Ligand identification based on enzymatic annotations

We extracted 911 unique gene and enzymatic reaction relationships, and 3785 small molecule and enzymatic reaction relationships from PlasmoCyc release 3.0 [26]. We mapped genes and the associated drug target scores to small molecules if they shared the same enzymatic reactions. Two-dimensional chemical structures of the small molecules represented by SMILE (Simplified Molecular Input Line Entry System) notation were also obtained from the same release. Three-dimensional structures of the ligands were extracted from PDB [74].

Computational screening of chemical databases using similarity of molecular fingerprints

Molecular fingerprints (bit-string representation of compound structures) were calculated using Pipeline Pilot program (Scitegic Inc., San Diego, CA) with element-class fingerprints with a diameter of 6 bonds (ECFP_6). Compound similarity was measured by the Tanimoto similarity coefficient of their fingerprints.

Two compound databases were used for the virtual screens. The bioactive compound database was composed of approximately 169,000 compounds including FDA approved drugs, drugs that are in the developmental phase and molecules that are used for large high throughput screens. The database of commercially available compounds was composed of approximately 1.5 million compounds from the following companies: ASINEX Inc, ChemBridge Inc, ChemDiv Inc, and Specs Inc.

Virtual screens were performed with Pipeline Pilot program. The first screen (against bioactives database) used a Tanimoto coefficient cutoff >=0.4. The second screen (against commercially available compound database) used a cutoff >=0.32.

ADME filters

ADME filters were composed of 1) Lipinski criteria: poor absorption or permeation is more likely when there are >5 H-bond donors, >10 H-bond acceptors, molecular weight >500 and the calculated Log P (CLogP) > 5 (or MlogP > 4.15) [38]; 2) Oprea criteria: "drug-like" compounds are more likely to be between the allowing limits: 0 <= hydrogen bond donors <= 2, 2 <=hydrogen bond acceptors<= 9, 2 <= the number of

rotatable bonds <= 8, and 1 <= the number of rings <4 [39]; 3) aqueous solubility predictions generated using Pipeline Pilot model (Scitegic Inc.); 4) REOS filters [40].

The ADME filtering procedure was applied to the compound set resulting from the second screen. Among the 7728 compounds, 3874 failed the solubility criteria (>10μg/ml), 88 compounds exceeded the maximum allowed number of violations using the combined Lipinski and Oprea criteria (violation <4), and 17 exceeded the REOS violations (<2).

RESULTS

Our informatics approach to identify antimalarial compounds involves the following five steps: (1) Generating phylogenomic profiles of the malaria proteome by comparing it to 203 completely sequenced genomes; (2) Mining the phylogenomic profiles for drug target genes by looking for genes that were conserved through evolution but lack significant mammalian homologs; (3) Identifying ligands that could potentially bind to those drug target proteins; (4) Screening large compound databases for drug-like compounds that are structurally similar to the ligands identified in step 3; and (5) testing antimalarial activity of candidate compounds by an *in vitro* growth inhibition assay (Figure 1).

Phylogenomic profiles of the *P. falciparum* proteome

202 completely sequenced genomes were used to construct the *P. falciparum* phylogenomic profile, including two mammalian genomes (*Homo sapiens*, *Rattus norvegicus*), 23 other eukaryotic genomes, 159 bacterial genomes and 18 archaea

genomes. Three apicomplexan parasitic species were included in this study including *P. falciparum*. The two additional species were *Plasmodium yoelii yoelii* (a causative agent for rodent malaria) [20] and *toxoplasma gondii* (an intracellular parasite that infects virtually all warm-blooded organisms and may cause disease and death in association with immunosuppressive conditions) [41].

The complete *P. falciparum* proteome sequence was blasted against all annotated protein sequences from the 202 genomes. For each malaria protein, the best pair-wise BLASTP score within each of the 202 genomes was extracted to form the phylogenetic profile of the *P. falciparum* protein. A matrix of the scores was compiled and clustered to form a phylogenomic profile of the *P. falciparum* proteome with each row representing a single *P. falciparum* protein and each column representing a genome (Figure 2).

In addition to sequence homology, several other metrics were used to construct phylogenomic profiles, which were sequence homology between orthologs (defined as the best reciprocal BLASTP hits between genomes), similarity to protein domain and family signatures defined in protein signature databases such as Pfam, TIGRFAMs, ProDom, SUPERFAMILY, SMART, PIRSF, PRINTS and InterPro [31,32,42-48]. Profiles were constructed in a similar way as those generated using sequence homology (Materials and Methods). Twelve phylogenomic profile matrixes were constructed, one for each comparison method. The sequence homology-based metrics produced *P. falciparum* centric profiles (each row represented a *P. falciparum* protein); while similarity of protein signature-based profiles did not (each row represented a single signature).

Figure 2 illustrates a sample phylogenomic profile constructed using sequence homology. The profile contained good phylogenetic signal on both the deep and shallow levels of the phylogeny scales. The three primary branches of the tree of life: Archaea, Eukaryota and Bacteria were well separated. Evolutionarily closely related genomes were clustered together such as the Enterobacteriales (Escherichia, Salmonella, Shigella and Yersinia), the Bacilli (Streptococcus, Lactococcus, Enterococcus, Listeria, Bacillus, Oceanobacillus and Staphylococcus) and the two closely related malaria species (*P. falciparum* and *P. yoelli yoelli*). Examining the clustering on the direction of the *P. falciparum* proteome, we could identify groups of genes with different phylogenetic profiles such as genes conserved through evolution and shared by most of the taxons (enolase *PF10_0155*); eukaryotic-specific genes such as proteins in the ubiquitin system (*MAL8P1.23, PF10_0330*); and genes conserved within archaea and eukaryotes, such as DNA replication mini-chromosome maintenance (MCM) proteins (*PFE1345c, PF14_0177, PF07_0023, PFL0580w, PF13_0095, PF13_0291*) [49]. An example of genes detected exclusively in the three apicomplexan parasitic genomes was trophozoite antigen R45 (*PFD1175w*) and its gene family [50]. Although most Apicomplexa-specific genes were annotated as "conserved hypothetical" proteins identified based on computational predictions, their conservation within Apicomplexa indicated their function might be related to the parasitic life style and evolved after the divergence of the common ancestor of apicomplexans. Most *P. falciparum* genes had homologs in the *P. yoelii yoelii* proteome, although 20% of its proteome were specific to the *P. falciparum* lineage. After examining the chromosomal locations of those *P. falciparum*-specific genes, they were shown to be predominantly located at the sub-telomeric regions. Many

of those *P. falciparum*-specific genes were members of the *var*, *rifin* and *stevor* gene families, which were expressed on infected erythrocyte membrane and were important factors possibly contributing to the parasite's antigenic variation [51-54]. These observations are consistent with the previous hypothesis that the sub-telomeric regions harbor fast evolving large gene families, indicating a possible functional role in antigenic variation and immune evasion. [20,54-56].

Mining the phylogenomic profiles for drug target genes

Our goal was to mine the phylogenomic profiles to identify the drug target genes - *P. falciparum* genes that are conserved through evolution but lacking significant mammalian homologs. They are potentially good targets because evolutionarily conserved genes are more likely to be essential, and lacking mammalian homologs meaning that they are less likely to cause adversarial effect in human. An added benefit was that conserved proteins were more likely to have been studied in other organisms. Information on protein function, structure and inhibitors of the homologous genes could assist our effort to find inhibitors for the malaria targets.

To identify genes with the above phylogenetic pattern, we developed a scoring function to estimate its conservation in non-mammalian genomes, panelized by its conservation in the mammalian genomes. The scoring function balanced the phylogenetic pattern generated using sequence homology and protein signatures by integrating information from all phylogenetic profile matrixes. For each *P. falciparum* gene, the drug target score (*S*) was calculated as the weighted sum of a term measuring the

conservation in non-mammalian genomes vs. mammalian genomes by the individual phylogenetic comparison method (Materials and Methods).

The distribution of the drug target score (*S*) is illustrated in Figure 3. The score ranges from -384 to +2739 with larger negative values meaning better drug targets. The average score is +48.29±119.19 (Supplemental Table S2). A sharp increase in the cumulative frequency was observed at approximately $S = -1.6$ with a majority (50%) of the proteins located between –1 and +10. To evaluate whether our scoring function was able to enrich for potential drug targets, we tested the scoring function on a list of 12 known targets compiled through a literature search [17,24,26,57-65], which have an average score –58± 83 (Supplemental Table S3). Ten of the 12 (83%) positive controls scored better than –1.6.

Using $S < -1.6$ as the threshold, 152 *P. falciparum* genes (3% of the proteome) were identified as candidate drug target genes (Table 1). The top 20 of the list is shown in Table 2. Among them was the entire *P. falciparum* isoprenoids biosynthesis pathway (three identified enzymes: 1-deoxy-D-xylulose-5-phosphate synthase, 1-deoxy-D-xylulose-5-phosphate reductoisomerase, 2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase [57,64,66] and four predicted enzymes: 2C-methyl-D-erythritol-4-phosphate cytidyltransferase, 4-diphophocytidyl-2c-methyl-D-erythritol kinase, GcpE, LytB [4,26]). Their phylogenetic profiles showed that the pathway was conserved in bacteria and plants, but was absent in vertebrates (Figure 4). This finding was consistent with previous studies that have identified this pathway as an excellent target for developing new anti-malarial drugs [57,64,67].

Identification of ligands bound to the drug target proteins

We identified ligands that could potentially bind to the drug target proteins using the following two methods: comparative protein structure modeling and enzymatic annotations.

Our first method was to identify co-binding ligands in protein structural models generated by comparative protein structure modeling. Comparative protein structure modeling, or homology modeling, is a computational method to predict three-dimensional protein structure models based primary on its alignment to known structures [25,68-72]. Structural models of high accuracy (in the range of 3Å) can be obtained when template structures and the modeled sequence share strong sequence homology (>50% sequence identity) [37,69]. Models for the *P. falciparum* proteome are available in MODBASE, a database of predicted structure models to all known protein sequences [36,37]. We identified 2019 *P. falciparum* proteins in MODBASE with at least one known and reliably predicted structure model (Materials and Methods)

If there was a co-crystallized ligand in the template structures, we assumed the ligands could potentially bind to the predicted models as well. Based on the above assumption, 704 unique ligands was identified from LIGBASE, a database comprising all ligand-binding sites of known protein structures in the Protein Data bank (PDB) [73] (Figure 5).

Our second method to predict binding ligands was to identify natural substrates and products of *P. falciparum* enzymes. Enzymatic annotations were obtained from PlasmoCyc, a pathway database for the *P. falciparum* genome including information on 737 enzymes, 816 enzymatic reactions, and 525 small molecules compounds [26,75-77].

335 unique small compounds were identified as natural substrates or products of a predicted enzyme in the *P. falciparum* genome (Materials and Methods) (Figure 5). Two-dimensional chemical structures of the small molecules were also obtained from PlasmoCyc.

The combined set of small molecules obtained using comparative structure modeling and enzymatic annotation was filtered to eliminate polymers, duplicates, molecules with < 5 atoms, and molecules that failed to convert to machine readable format, which yielded 780 ligands.

Drug target scores for *P. falciparum* proteins was subsequently transferred to their corresponding ligands. In some cases, a single ligand was associated with multiple proteins, which yielded multiple drug target scores. To capture the entire range of scores, we assigned maximum and minimum scores to each ligand. Using the same cutoff −1.6 as that used for identifying drug target proteins, we obtained 56 ligands with both maximum and minimum scores passing the threshold and 21 ligands with only the minimum score passing the threshold. In total, the phylogenetic filter yielded 77 ligands (Figure 5 and Supplemental Table S4).

Screening large chemical databases for drug-like compounds

In addition to binding to malaria target proteins and causing low toxicity to human cells, successful antimalarials also possess other drug-like properties such as good bioactivity and favorable bioavailability in the human body. Studies have shown poor pharmacokinetics is an important cause of failure in drug development and should be analyzed as early as possible in the drug discovery process. To facilitate screening out

compounds with unfavorable bioavailability, computation models for drug absorption, distribution, metabolism and excretion (ADME) properties have been developed [28,79,80]. In addition, we desired compounds that could be purchased commercially for efficient experimental validation.

A series of computational screening and modeling procedures was carried out to find compounds that were structurally similar to our ligands and optimized for the other properties (Figure 5).

To improve bioactivity of our compounds, we computationally screened a large database of bioactives (169 K) to identify compounds with documented bioactivity. Ligand structural similarity was measured by Tanimoto similarity coefficient of bit-string representation of molecular structures (molecular fingerprints) [78]. This screen yielded 728 bioactive compounds. Using the 728 compounds from the first screen, a second screen of a database comprised of commercially available compounds (1.5 M) yielded 7728 compounds. The set of 7728 compounds was subsequently subjected to an ADME filtering procedure, which yielded 3749 good compounds (Materials and Methods). Upon examining the distribution of the 3749 compounds by their initial 77 query ligands, certain ligands dominated the selection. To increase the diversity in the final compound set, the maximum number of compounds per query ligand was capped at 125 per commercial vendor, which yielded 1893 compounds from four commercial vendors (300 - 600 if purchased from a single company) (Supplemental Figure S1).

The screening process could best be illustrated by an example of a query ligand, such as *p*-aminobenzoate. *p*-Aminobenzoate was initially identified a natural substrate or product of dihydropteroate synthetase and para-aminobenzoic acid synthetase. Both

enzymes were among the list of candidate drug target genes (dihydropteroate synthetase with a drug target score -36 and para-aminobenzoic acid synthetase scored -68). The drug target scores were transferred from the enzymes to *p*-aminobenzoate (minimum score: –68 and maximum score: –36). Since both its minimum and maximum scores passed the –1.6 threshold, *p*-aminobenzoate was selected as a query ligand to proceed to the two-step virtual screen process. In the first step, *p*-aminobenzoate was screened against compounds in the bioactive database, which yielded 31 structurally similar compounds including *p*-aminobenzoate. In the second step, the 31 compounds were screened against compounds in four commercial collections. After filtering out compounds with unfavorable ADME properties and applying the diversity filter, 125 compounds were derived from each commercial vendor database (Figure 6).

DISCUSSION

Experimental validation

The experimental validation step of this study is currently underway. 646 compounds were ordered from ChemBridge Inc. (San Diego, CA) (Supplemental Table S5). The antimalarial activity of those compounds will be tested using a high-throughput *in-vitro* growth inhibition assay as described in previous studies [81,82]. The same experimental assay will be applied on a negative control set, composed of an equal number of compounds chosen randomly from an in-house collection. The effectiveness of this informatics approach will be evaluated by the enrichment of antimalarial activity in selected compounds compared to the negative control.

The ideal negative control sets

In addition to enriching for antimalarials, our approach was also aimed at selecting compounds that will cause lower toxicity in human cells by targeting *P. falciparum* genes without significant homologs in mammalian genomes. The toxicity of our candidate compounds could be measured in a human cell line and evaluated by comparing to a second negative control set composed of compounds selected through the same informatics procedure except their target genes have a high drug target score (opposite pattern of phylogenetic profiles: conserved in mammalian genomes without significant homologs in non-mammalian genomes).

Therefore, the ideal negative control should include two sets of compounds. The first set would be the randomly selected compounds for the evaluation of antimalarial activities. The second negative control set for the evaluation of toxicity in a human cell line. Both sets should also be filtered through the same ADME and diversity filters and ordered from the same vendor to eliminate any bias introduced by those procedures.

The two ideal negative control sets were generated computationally (Supplemental Table S6). Due to financial limitations, we did not purchase them; instead we used an in-house collection as our only negative control. We were therefore unable to evaluate the human toxicity aspect experimentally. In addition, due to lack of control over drug-like properties, the antimalarial activity of the negative control set will have certain level of bias, depending on the specific compounds that will be chosen.

Approximation in the computational process

Approximation, simplification and their associated errors are unavoidable components of any computational approach for drug development. Several significant assumptions and approximations were applied in our computational process.

In the process of drug target identification

In the process of using a phylogenomic profile to identify drug target genes, we used conservation through evolution as a proxy for gene essentiality. To find out whether a gene is essential, ideally we need to knock out (or conditionally knock out) the gene product experimentally. Unfortunately, the genome-wide deletion project in *Plasmodium* has not been carried out. Alternative computational approaches to assign essentiality included mapping essential genes identified in model organisms such as *E. coli* and *S. cerevisiae* to *P. falciparum* genome [83], identifying components that lack alternative paths in metabolic networks [26,84], and identifying "network hubs" in protein interaction networks [85,86]. In addition, we used the lack of significant mammalian homologs as the proxy to low toxicity. The caveat of this assumption was that it did not consider small molecule cross-reaction to other protein targets. To identify orthologs, we simply used reciprocal inter-genome best BLASTP hits as biological orthologs without running full phylogenetic reconstruction. Although more accurate computational methods were available [87,88], we did not expect they would dramatically change our drug target gene list since the contribution from ortholog analysis only accounted for a fraction (1/12) of the final results. In addition to those assumptions, the accuracy of our phylogenomic analysis was also dependent on the performance of sequence homology search and protein signature identification.

In the process of predicting binding ligands

Additional uncertainty was introduced when we used protein structure and co-ligand predictions available in MODBASE and LIGBASE. Comparative structure modeling can achieve relatively high accuracy (in the range of 3Å) when template structures and the modeled sequence share strong sequence homology (>50% sequence identity). Sequence alignments on which the models are based generally contain almost no errors. Models of this level of accuracy can be used for docking of small ligands [89]. If the sequence identity is in the range of 30–50%, models tend to have >85% of the C atoms within 3.5 Å of their correct positions [37,69]. Models of this range of accuracy correspond to the reliability score 0.7, which can be used to predict the location of the binding site, but not enough for docking small molecules [69]. We used 0.7 as the cutoff for the structure predictions, which meant that a subset of models did not have enough accuracy for directly predicting co-binding ligands.

An even greater simplification was made when we assumed co-crystallized ligands in structure templates would bind to the predicted structure models as well (also the premise of LIGBASE). A more thorough computational approach should compare the binding sites (under development in LIGBASE, unpublished data) and perform molecular docking of the ligand to the structure models. These improvements required either further development of computational methods or a large amount of manual intervention that were beyond our available resources. With the approximation in both structure prediction and transferring of co-ligands, not all predicted binding ligands would bind to the presumed *P. falciparum* drug targets.

Comparing to structure prediction-based binding ligand identification, we expected a lower error rate in the method based on natural substrates of enzymes, because

reaction chemistry, substrate specificity and the structure of the active sites are more conserved than allover structures and sequences [90-94]. Errors could still exist due to mistakes in ORF annotations (assigning ORFs as specific enzymes), the existence of certain *P. falciparum* pathways, and alternative substrates used by a *P. falciparum* enzyme.

In the process of virtual screens

Virtual screens of chemical databases are based on the idea that significant similarities in molecular structures are attributed to similarity in biological activities [95]. But structure and activity can relate in so many different ways that it is difficult to capture in bit string representation of molecular structures. Using a sequential two-step screening procedure with similarity score cutoffs 0.4 and 0.32, we expected a certain level of false positives and false negatives generated from the screen.

In addition to the similarity search, we implemented several rule-based methods (Lipinski, Oprea, and REOS rules) to predict "drug-like" properties. These rules were mostly generated from statistical analysis of structure activity relationships of various collections of compounds to classify them as "drug-like" or not "drug-like"; each rule generating somewhat different classifications and with different self-claimed recall rates for "drug-like" compounds (90% for Lipinski's "rule of 5"; 70% for Oprea criteria) [38-40]. We used a combination of Lipinski, Oprea, and REOS rules to maximize our chance to eliminate compounds with unfavorable pharmacokinetics properties. However, we still expected a subset of our compounds would not be perfect.

Using enrichment of antimalarial activities as the benchmark for evaluation

These approximations and simplifications introduced errors throughout the process. In addition, many other factors complicates the *in vitro* growth inhibition assay (used as an approximation for *in vivo* antimalarial activity), such as the ability of the compounds to cross the red blood cell membrane. With these uncertainties difficult to account for computationally, a majority of our compounds were not expected to come out as strong antimalarials. However, this computational approach could still increase the possibility of promising compounds. Enrichment by an informatics approach can concentrate future drug development resources on a more promising set. The hit rate for an experimental high-throughput screening (HTS) of compound libraries (without selection) is often quite low, typically well below 1% [96-98]. Studies have shown that a virtual screen can improve the hit rate by 20 fold to 1700 fold [96,97]. To demonstrate the utility of our approach, we will use the enrichment factor (hit rate of selected compounds divided by that of the negative control) instead of absolute antimalarial activities for evaluation.

Summary

The most important contribution of this manuscript is the development and implementation of this virtual drug development framework to discover antimalarial compounds *in silico*. This framework started with 203 complete genome sequences and resulted in 1893 potential antimalarials; in the process it integrated a diverse and large amount of informatics dataset of protein signatures and profiles, metabolic pathways, protein 3D structure models, and large compound collections. A large spectrum of computational methods was used or developed in this framework, which included

sequence homology searches, ortholog identification, a phylogenomic analysis of a complete proteome, prediction for drug-like compounds, molecular fingerprints, a scoring function to systematically identify drug target proteins from the complete *P. falciparum* genome, and a virtual screen procedure composed of a two-step similarity search followed by ADME and diversity filtering procedures. Although most of the informatics and computational components have been developed, they were rarely put together in a single pipeline. Even more unusual, the derived compounds can now be tested experimentally.

Although uncertainties and errors were associated with almost every computational step, these computational procedures made this framework extremely high-throughput, much cheaper to perform than a pure experimental screening approach, and capable of utilizing the tremendous amount of information accumulated in sequencing projects, functional genomics, structural genomics, and computational chemistry.

We have demonstrated the utility of this framework to develop inhibitors against one of the most deadly human parasites in the world. Other infectious diseases, many of which are associated with poverty and neglect, can benefit from the same pipeline as well.

Reference

1. WHO (2005) World malaria report 2005.

2. WHO (2002) Communicable diseases.

3. Yamey G (2004) Roll Back Malaria: a failing global health campaign. Bmj 328: 1086-1087.

4. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498-511.

5. Korenromp E (2005) MALARIA INCIDENCE ESTIMATES AT COUNTRY LEVEL FOR THE YEAR 2004 – PROPOSED ESTIMATES AND DRAFT REPORT -. In: Roll Back Malaria WHO, editor: World Health Organization.

6. Bray PG, Ward SA, O'Neill PM (2005) Quinolines and artemisinin: chemistry, biology and history. Curr Top Microbiol Immunol 295: 3-38.

7. Bloland PB (2001) Drug resistance in malaria. In: WHO/CDS/CSR/DRS/2001.4, editor.

8. Jefford CW (2001) Why artemisinin and certain synthetic peroxides are potent antimalarials. Implications for the mode of action. Curr Med Chem 8: 1803-1826.

9. Jung M, Lee K, Kim H, Park M (2004) Recent advances in artemisinin and its derivatives as antimalarial and antitumor agents. Curr Med Chem 11: 1265-1284.

10. Talisuna AO, Bloland P, D'Alessandro U (2004) History, dynamics, and public health importance of malaria parasite resistance. Clin Microbiol Rev 17: 235-254.

11. Nabarro DN, Tayler EM (1998) The "roll back malaria" campaign. Science 280: 2067-2068.

12. D'Alessandro U, Buttiens H (2001) History and importance of antimalarial drug resistance. Trop Med Int Health 6: 845-848.

13. Looareesuwan S, Buchachart K, Wilairatana P, Chalermrut K, Rattanapong Y, et al. (1997) Primaquine-tolerant vivax malaria in Thailand. Ann Trop Med Parasitol 91: 939-943.

14. Hamedi Y, Nateghpour M, Soonthornsata B, Tan-ariya P, Kojima S, et al. (2003) Monitoring of Plasmodium vivax sensitivity to chloroquine in vitro in Thailand. Trans R Soc Trop Med Hyg 97: 435-437.

15. Murphy GS, Basri H, Purnomo, Andersen EM, Bangs MJ, et al. (1993) Vivax malaria resistant to treatment and prophylaxis with chloroquine. Lancet 341: 96-100.

16. Khan MA, Smego RA, Jr., Razi ST, Beg MA (2004) Emerging drug--resistance and guidelines for treatment of malaria. J Coll Physicians Surg Pak 14: 319-324.

17. Surolia N, Surolia A (2001) Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of Plasmodium falciparum. Nat Med 7: 167-173.

18. Ralph SA, Van Dooren GG, Waller RF, Crawford MJ, Fraunholz MJ, et al. (2004) Tropical infectious diseases: Metabolic maps and functions of the Plasmodium falciparum apicoplast. Nat Rev Microbiol 2: 203-216.

19. Ralph SA, D'Ombrain MC, McFadden GI (2001) The apicoplast as an antimalarial drug target. Drug Resist Updat 4: 145-151.

20. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, et al. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. Nature 419: 512-519.

21. Bahl A, Brunk B, Crabtree J, Fraunholz MJ, Gajria B, et al. (2003) PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. Nucleic Acids Res 31: 212-215.

22. Koonin EV, Galperin MY (2003) Sequence - evolution - function: computational approaches in comparative genomics. Boston: Kluwer Academic. xiii, 461 p., [411] p. of plates p.

23. Izaki K, Matsuhashi M, Strominger JL (1968) Biosynthesis of the peptidoglycan of bacterial cell walls. 8. Peptidoglycan transpeptidase and D-alanine carboxypeptidase: penicillin-sensitive enzymatic reaction in strains of Escherichia coli. J Biol Chem 243: 3180-3192.

24. Waller RF, Ralph SA, Reed MB, Su V, Douglas JD, et al. (2003) A type II pathway for fatty acid biosynthesis presents drug targets in Plasmodium falciparum. Antimicrob Agents Chemother 47: 297-301.

25. Fiser A, Sali A (2003) Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 374: 461-491.

26. Yeh I, Hanekamp T, Tsoka S, Karp PD, Altman RB (2004) Computational analysis of Plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. Genome Res 14: 917-924.

27. Ewing TJ, Makino S, Skillman AG, Kuntz ID (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 15: 411-428.

28. van de Waterbeemd H, Gifford E (2003) ADMET in silico modelling: towards prediction paradise? Nat Rev Drug Discov 2: 192-204.

29. Miller MA (2002) Chemical database techniques in drug discovery. Nat Rev Drug Discov 1: 220-227.

30. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

31. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2005) InterPro, progress and status in 2005. Nucleic Acids Res 33 Database Issue: D201-205.

32. Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families. Nucleic Acids Res 31: 371-373.

33. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, et al. (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res 33: D154-159.

34. Zdobnov EM, Apweiler R (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847-848.

35. Melo F, Sanchez R, Sali A (2002) Statistical potentials for fold assessment. Protein Sci 11: 430-448.

36. Pieper U, Eswar N, Stuart AC, Ilyin VA, Sali A (2002) MODBASE, a database of annotated comparative protein structure models. Nucleic Acids Res 30: 255-259.

37. Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, et al. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 32: D217-222.

38. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46: 3-26.

39. Oprea TI (2000) Property distribution of drug-related chemical databases. J Comput Aided Mol Des 14: 251-264.

40. Walters WP, Murcko MA (2002) Prediction of 'drug-likeness'. Adv Drug Deliv Rev 54: 255-271.

41. Kissinger JC, Gajria B, Li L, Paulsen IT, Roos DS (2003) ToxoDB: accessing the Toxoplasma gondii genome. Nucleic Acids Res 31: 234-236.

42. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29: 37-40.

43. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Res 32: D138-141.

44. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, et al. (2005) The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res 33 Database Issue: D212-215.

45. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, et al. (2003) PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 31: 400-402.

46. Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, et al. (2004) PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res 32: D112-114.

47. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, et al. (2004) SMART 4.0: towards genomic data integration. Nucleic Acids Res 32: D142-144.

48. Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J (2004) The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Res 32: D235-239.

49. Chong JP, Hayashi MK, Simon MN, Xu RM, Stillman B (2000) A double-hexamer archaeal minichromosome maintenance protein is an ATP-dependent DNA helicase. Proc Natl Acad Sci U S A 97: 1530-1535.

50. Schneider AG, Mercereau-Puijalon O (2005) A new Apicomplexa-specific protein kinase family: multiple members in Plasmodium falciparum, all with an export signature. BMC Genomics 6: 30.

51. Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, et al. (1995) Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. Cell 82: 101-110.

52. Cheng Q, Cloonan N, Fischer K, Thompson J, Waine G, et al. (1998) stevor and rif are Plasmodium falciparum multicopy gene families which potentially encode variant antigens. Mol Biochem Parasitol 97: 161-176.

53. Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, et al. (1995) The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. Cell 82: 89-100.

54. Janssen CS, Phillips RS, Turner CM, Barrett MP (2004) Plasmodium interspersed repeats: the major multigene superfamily of malaria parasites. Nucleic Acids Res 32: 5712-5720.

55. Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, et al. (2005) A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. Science 307: 82-86.

56. Carlton J, Silva J, Hall N (2005) The genome of model malaria parasites, and comparative genomics. Curr Issues Mol Biol 7: 23-37.

57. Jomaa H, Wiesner J, Sanderbrand S, Altincicek B, Weidemeyer C, et al. (1999) Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. Science 285: 1573-1576.

58. Triglia T, Cowman AF (1994) Primary structure and expression of the dihydropteroate synthetase gene of Plasmodium falciparum. Proc Natl Acad Sci U S A 91: 7149-7153.

59. Sixsmith DG, Watkins WM, Chulay JD, Spencer HC (1984) In vitro antimalarial activity of tetrahydrofolate dehydrogenase inhibitors. Am J Trop Med Hyg 33: 772-776.

60. Wiesner J, Sanderbrand S, Altincicek B, Beck E, Jomaa H (2001) Seeking new targets for antiparasitic agents. Trends Parasitol 17: 7-8.

61. McConkey GA, Pinney JW, Westhead DR, Plueckhahn K, Fitzpatrick TB, et al. (2004) Annotating the Plasmodium genome and the enigma of the shikimate pathway. Trends Parasitol 20: 60-65.

62. McRobert L, McConkey GA (2002) RNA interference (RNAi) inhibits growth of Plasmodium falciparum. Mol Biochem Parasitol 119: 273-278.

63. Lichtenthaler HK (2000) Non-mevalonate isoprenoid biosynthesis: enzymes, genes and inhibitors. Biochem Soc Trans 28: 785-789.

64. Kemp LE, Bond CS, Hunter WN (2002) Structure of 2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase: an essential enzyme for isoprenoid biosynthesis and target for antimicrobial drug development. Proc Natl Acad Sci U S A 99: 6591-6596.

65. Perozzo R, Kuo M, Sidhu AS, Valiyaveettil JT, Bittman R, et al. (2002) Structural elucidation of the specificity of the antibacterial agent triclosan for malarial enoyl acyl carrier protein reductase. J Biol Chem 277: 13106-13114.

66. Rohdich F, Eisenreich W, Wungsintaweekul J, Hecht S, Schuhr CA, et al. (2001) Biosynthesis of terpenoids. 2C-Methyl-D-erythritol 2,4-cyclodiphosphate synthase (IspF) from Plasmodium falciparum. Eur J Biochem 268: 3190-3197.

67. Lell B, Ruangweerayut R, Wiesner J, Missinou MA, Schindler A, et al. (2003) Fosmidomycin, a novel chemotherapeutic agent for malaria. Antimicrob Agents Chemother 47: 735-738.

68. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294: 93-96.

69. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, et al. (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29: 291-325.

70. Sanchez R, Sali A (2000) Comparative protein structure modeling. Introduction and practical examples with modeller. Methods Mol Biol 143: 97-129.

71. Sanchez R, Sali A (1997) Evaluation of comparative protein structure modeling by MODELLER-3. Proteins Suppl 1: 50-58.

72. Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M (1995) Evaluation of comparative protein modeling by MODELLER. Proteins 23: 318-326.

73. Stuart AC, Ilyin VA, Sali A (2002) LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. Bioinformatics 18: 200-201.

74. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.

75. Karp PD, Riley M, Paley SM, Pellegrini-Toole A (2002) The MetaCyc Database. Nucleic Acids Res 30: 59-61.

76. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, et al. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 32: D438-442.

77. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, et al. (2000) The EcoCyc and MetaCyc databases. Nucleic Acids Res 28: 56-59.

78. Willett P (2000) Chemoinformatics - similarity and diversity in chemical libraries. Curr Opin Biotechnol 11: 85-88.

79. Yamashita F, Hashida M (2004) In silico approaches for predicting ADME properties of drugs. Drug Metab Pharmacokinet 19: 327-338.

80. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, et al. (2002) Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 45: 2615-2623.

81. Madrid PB, Wilson NT, DeRisi JL, Guy RK (2004) Parallel synthesis and antimalarial screening of a 4-aminoquinoline library. J Comb Chem 6: 437-442.

82. Anderson MO, Sherrill J, Madrid PB, Liou AP, Weisman JL, et al. (2005) Parallel synthesis of 9-aminoacridines and their evaluation against chloroquine-resistant Plasmodium falciparum. Bioorg Med Chem.

83. Zhang R, Ou HY, Zhang CT (2004) DEG: a database of essential genes. Nucleic Acids Res 32: D271-272.

84. Palumbo MC, Colosimo A, Giuliani A, Farina L (2005) Functional essentiality from topology features in metabolic networks: a case study in yeast. FEBS Lett 579: 4642-4646.

85. Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M (2004) Genomic analysis of essentiality within protein networks. Trends Genet 20: 227-231.

86. Coulomb S, Bauer M, Bernard D, Marsolier-Kergoat MC (2005) Gene essentiality and the topology of protein interaction networks. Proc Biol Sci 272: 1721-1725.

87. Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189.

88. O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33: D476-480.

89. Ring CS, Sun E, McKerrow JH, Lee GK, Rosenthal PJ, et al. (1993) Structure-based inhibitor design by using protein models for the development of antiparasitic agents. Proc Natl Acad Sci U S A 90: 3583-3587.

90. Sanchez R, Sali A (1998) Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc Natl Acad Sci U S A 95: 13597-13602.

91. Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307: 1113-1143.

92. Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41: 98-107.

93. Wilson CA, Kreychman J, Gerstein M (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 297: 233-249.

94. Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, et al. (2001) The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. J Mol Biol 311: 693-708.

95. Johnson MA, Maggiora GM, American Chemical Society. Meeting (1990) Concepts and applications of molecular similarity. New York: Wiley. xix, 393 p. p.

96. Paiva AM, Vanderwall DE, Blanchard JS, Kozarich JW, Williamson JM, et al. (2001) Inhibitors of dihydrodipicolinate reductase, a key enzyme of the diaminopimelate pathway of Mycobacterium tuberculosis. Biochim Biophys Acta 1545: 67-77.

97. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, et al. (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem 45: 2213-2221.

98. Polgar T, Baki A, Szendrei GI, Keseru GM (2005) Comparative virtual and experimental high-throughput screening for glycogen synthase kinase-3beta inhibitors. J Med Chem 48: 7946-7959.

Figure 5-1 An *in silico* framework for antimalarial drug development

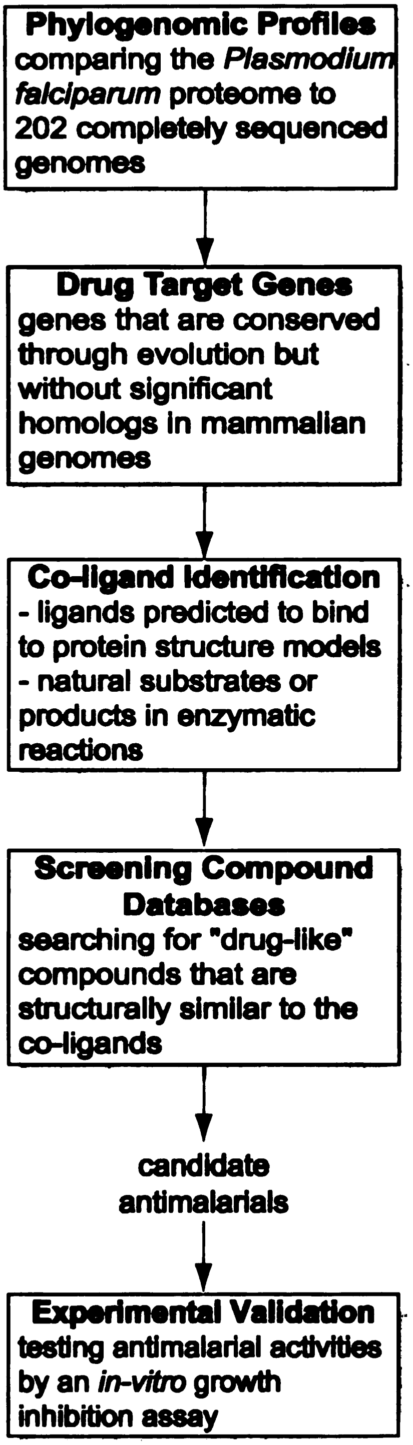
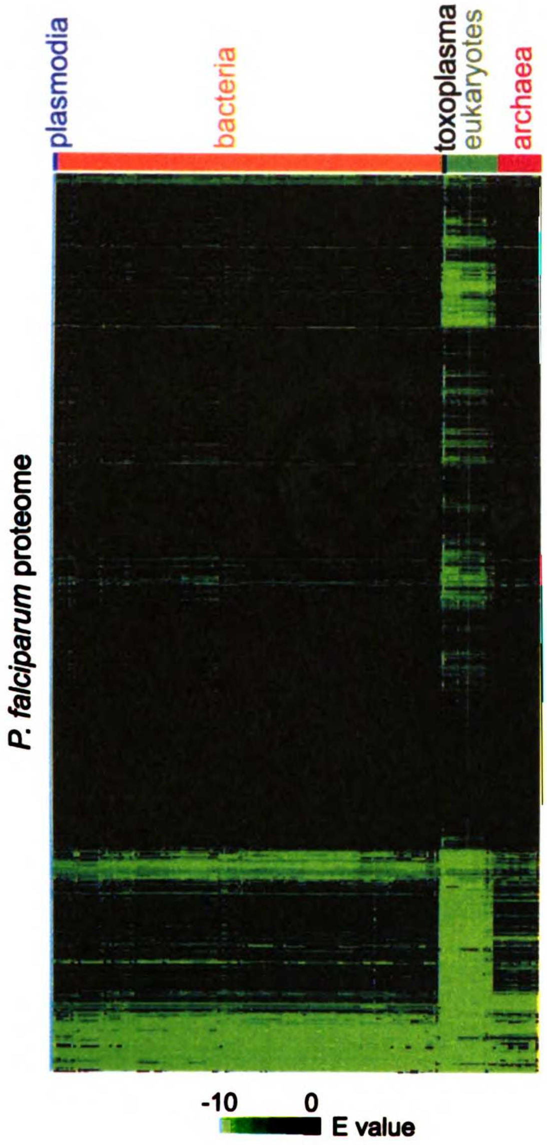

Phylogenomic Profiles
comparing the Plasmodium falciparum proteome to 202 completely sequenced genomes
Drug Target Genes
genes that are conserved through evolution but without significant homologs in mammalian genomes
Co-ligand Identification
- ligands predicted to bind to protein structure models
- natural substrates or products in enzymatic reactions
Screening Compound Databases
searching for "drug-like" compounds that are structurally similar to the co-ligands
candidate antimalarials
Experimental Validation
testing antimalarial activities by an in-vitro growth inhibition assay

Figure 5-2 A Phylogenomic profile of *Plasmodium falciparum*



eukaryotic specific genes
proteins in the ubiquitin system (*MAL8P1.23, PF10_0330*)
histone H2A (*MAL6P1.249*)
zinc finger protein (*PF13_0313*)
ras family GTPase, putative (*PFI0155c*)

Apicomplexa specific genes
trophozoite antigen R45 (*PFD1175w*)
conserved hypothetical proteins
erythrocyte membrane-associated antigen, putative (*PFD1045c*)

***P. falciparum* specific genes**
var, rifin, stevor

genes conserved within archaea and eukaryotes
Mini-Chromosome Maintenance (MCM) proteins (*PFE1345c, PF14_0177, PF07_0023, PFL0580w*)
flap exonuclease, putative (*PFD0420c*)

genes conserved through evolution
enolase (*PF10_0155*)
cysteine, isoleucine tRNA ligases, putative (*PF10_0149, PF13_0179*)

Figure 5-3 Distribution of the drug target score (*S*) in the *P. falciparum* proteome

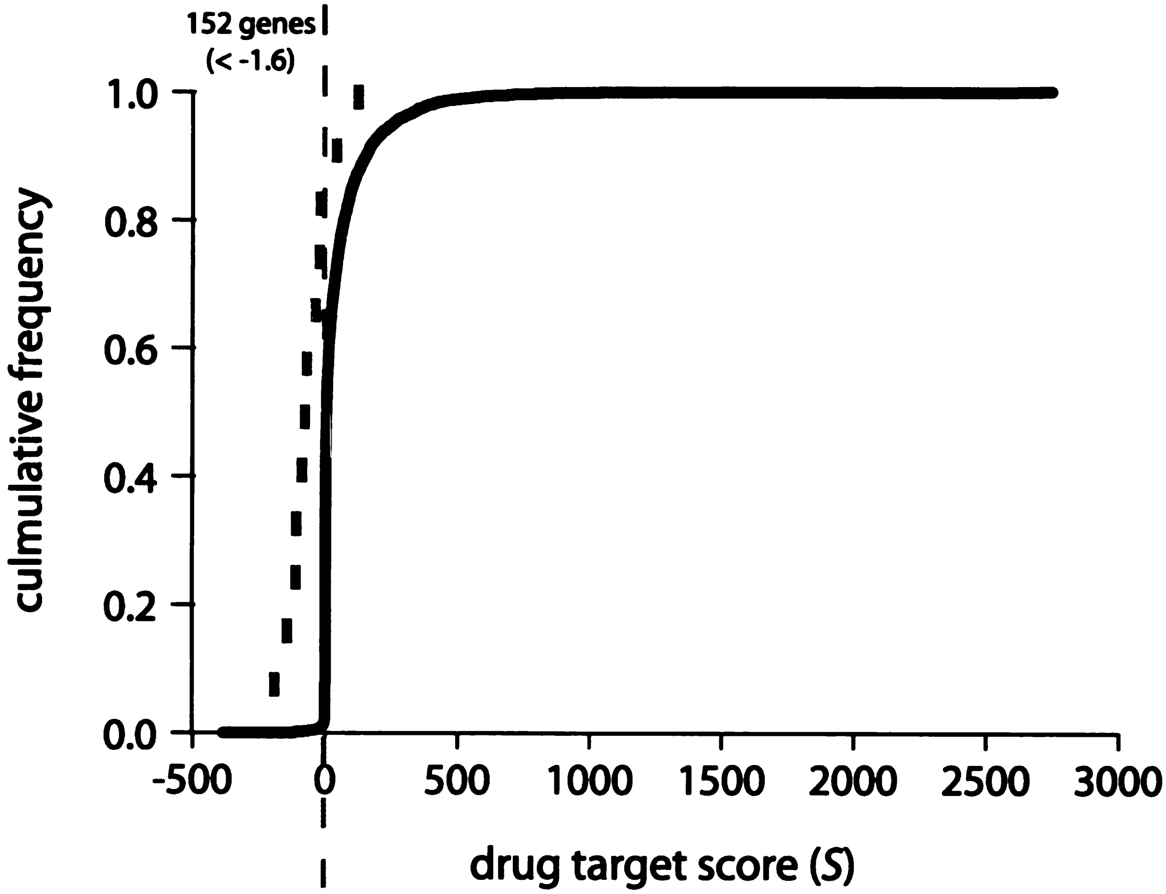
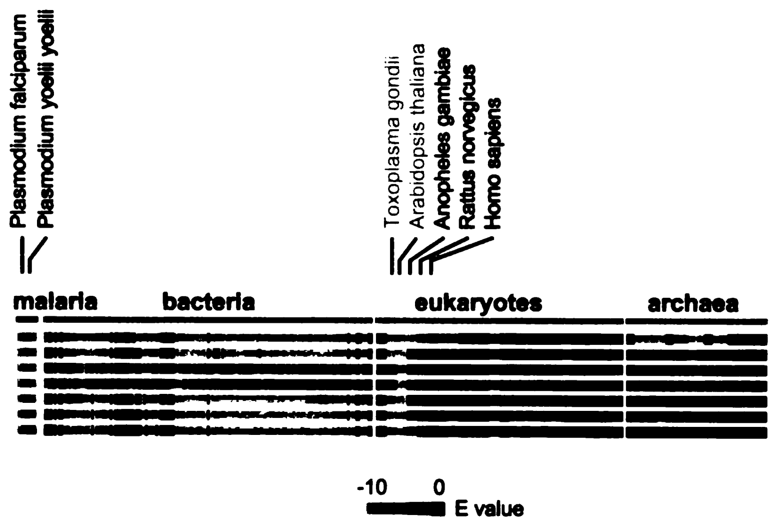


Figure 5-4 The phylogenetic profile and drug target scores of the isoprenoids biosynthesis pathway



	Drug Target Score (*S*)
1-deoxy-D-xylulose 5-phosphate synthase	-22.0
1-deoxy-D-xylulose 5-phosphate reductoisomerase	-190.2
2C-methyl-D-erythritol-4-phosphate cytidyltransferase	-3.2
4-diphophocytidyl-2c-methyl-D-erythritol kinase	-2.4
2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	-108.0
GcpE	-69.0
LytB	-143.3

Figure 5-5 Flowchart of the virtual drug development process for antimalarials

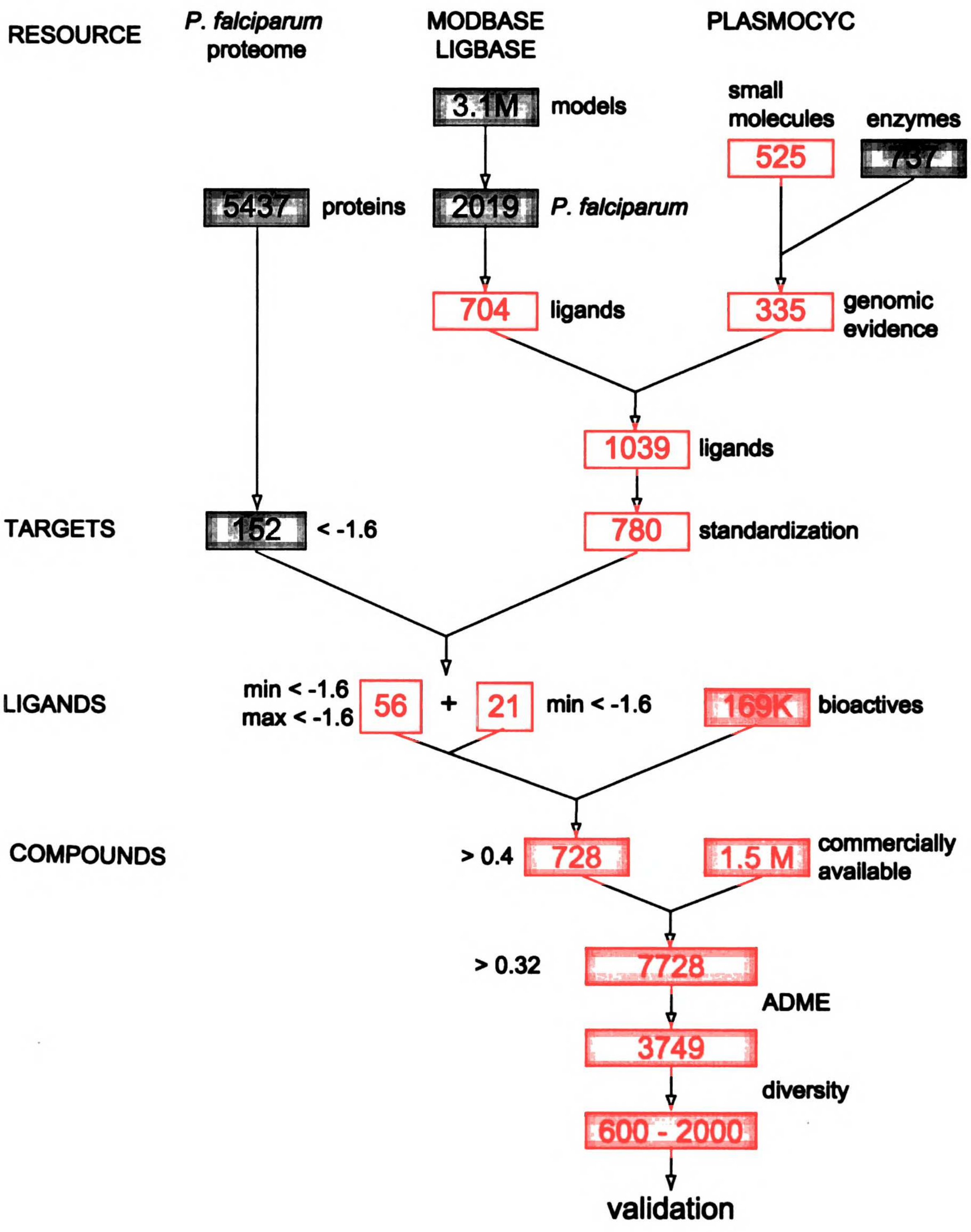


Figure 5-6 The identification of compounds that are structurally similar to *p*-aminobenzoate as potential antimalarials

p-Aminobenzoate was identified as a natural substrate or product for two drug target genes (dihydropteroate synthetase and para-aminobenzoic acid synthetase). Compounds that are structurally similar to *p*-aminobenzoate were identified using the two-step virtual screen procedure. A subset of the resulting compounds was shown.

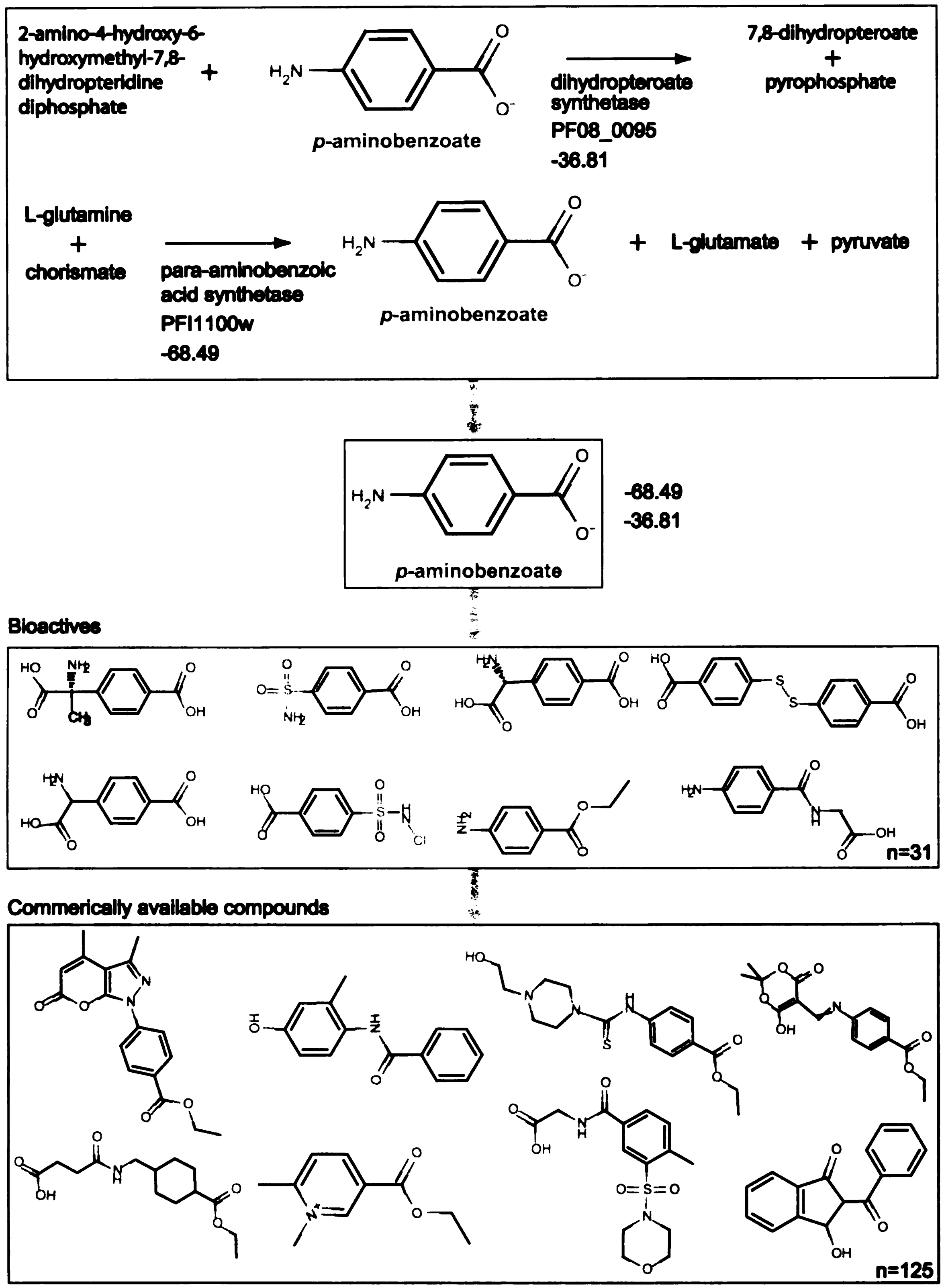

2-amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine diphosphate
+
p-aminobenzoate
dihydropteroate synthetase
PF08_0095
-36.81
7,8-dihydropteroate
+
pyrophosphate
L-glutamine
+
chorismate
para-aminobenzoic acid synthetase
PFI1100w
-68.49
p-aminobenzoate
+ L-glutamate + pyruvate
p-aminobenzoate
-68.49
-36.81
Bioactives
n=31
Commerically available compounds
n=125

Figure 5-S1 Compound frequency in the final selected set (per query ligand)

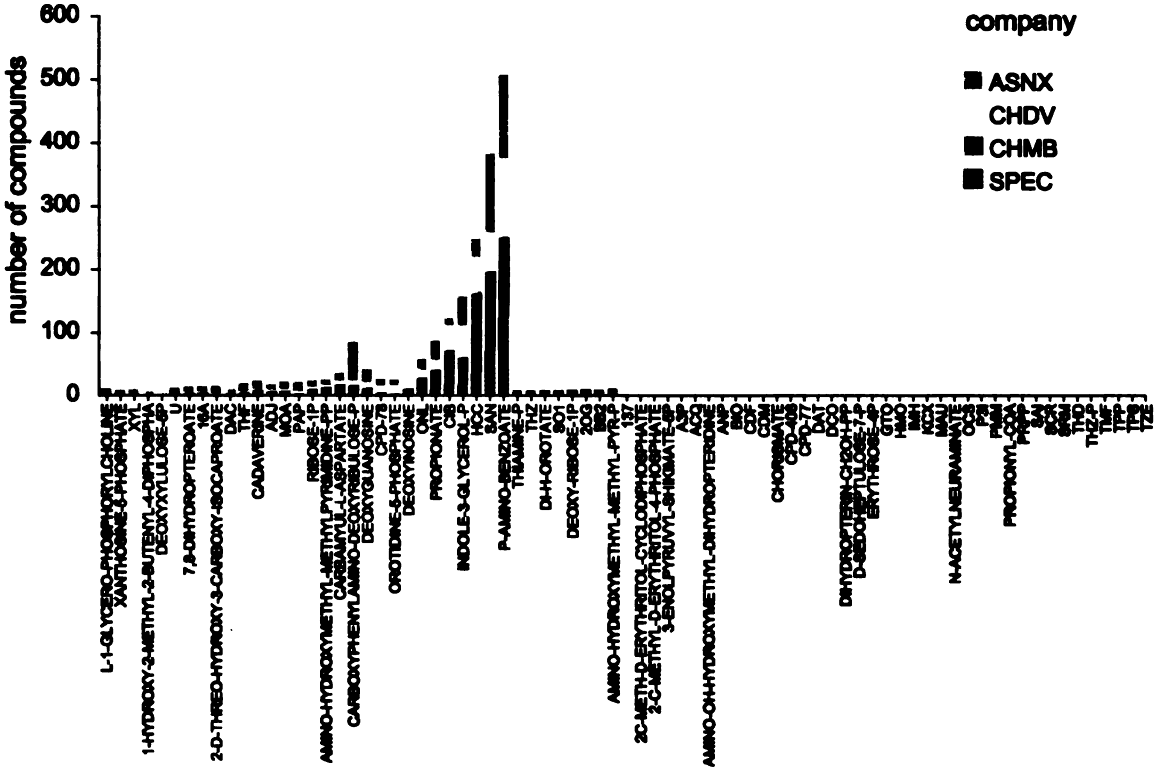
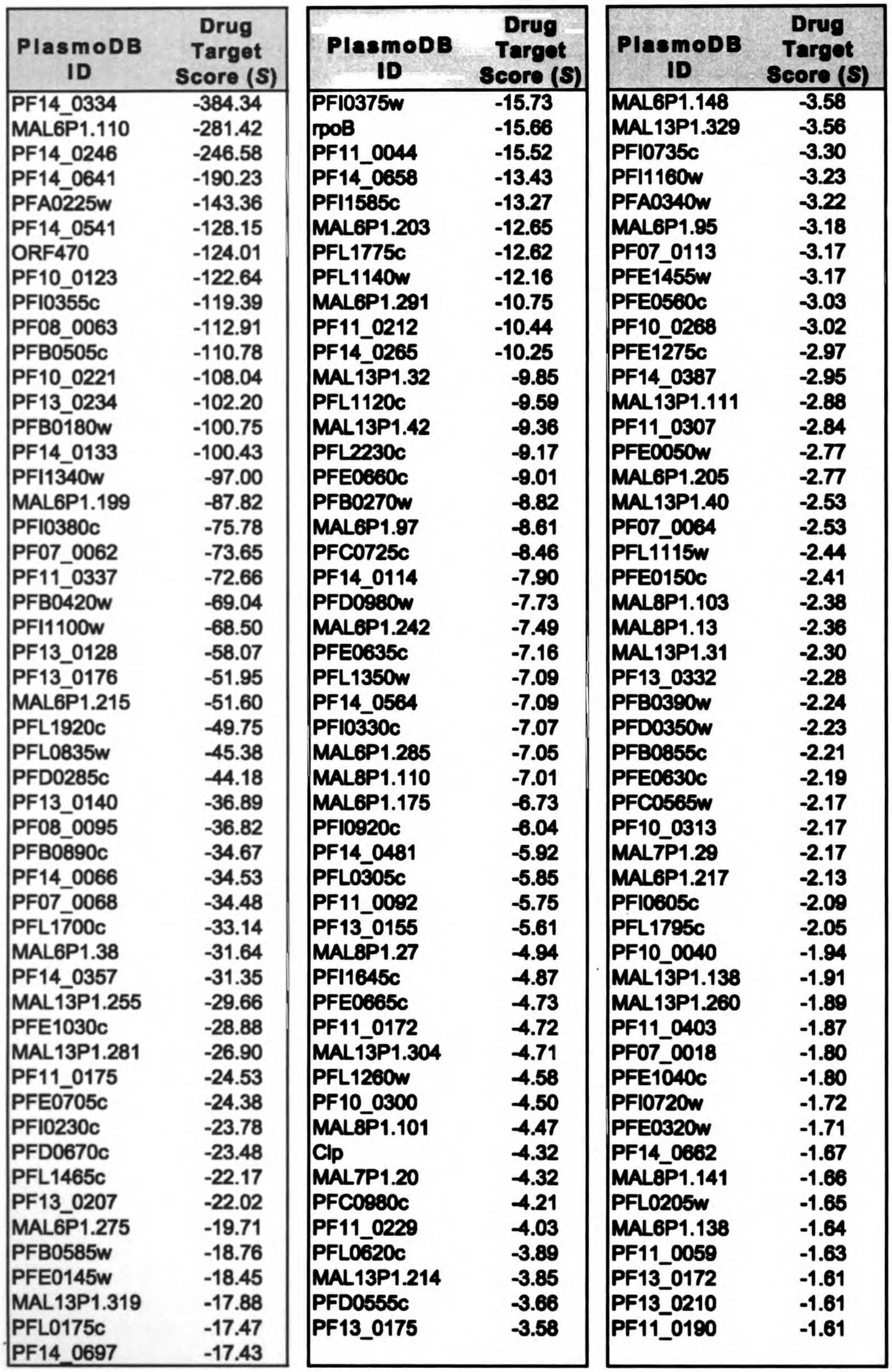


Table 5-1 Candidate drug target genes



PlasmoDB ID	Drug Target Score (*S*)
PF14_0334	-384.34
MAL6P1.110	-281.42
PF14_0246	-246.58
PF14_0641	-190.23
PFA0225w	-143.36
PF14_0541	-128.15
ORF470	-124.01
PF10_0123	-122.64
PFI0355c	-119.39
PF08_0063	-112.91
PFB0505c	-110.78
PF10_0221	-108.04
PF13_0234	-102.20
PFB0180w	-100.75
PF14_0133	-100.43
PFI1340w	-97.00
MAL6P1.199	-87.82
PFI0380c	-75.78
PF07_0062	-73.65
PF11_0337	-72.66
PFB0420w	-69.04
PFI1100w	-68.50
PF13_0128	-58.07
PF13_0176	-51.95
MAL6P1.215	-51.60
PFL1920c	-49.75
PFL0835w	-45.38
PFD0285c	-44.18
PF13_0140	-36.89
PF08_0095	-36.82
PFB0890c	-34.67
PF14_0066	-34.53
PF07_0068	-34.48
PFL1700c	-33.14
MAL6P1.38	-31.64
PF14_0357	-31.35
MAL13P1.255	-29.66
PFE1030c	-28.88
MAL13P1.281	-26.90
PF11_0175	-24.53
PFE0705c	-24.38
PFI0230c	-23.78
PFD0670c	-23.48
PFL1465c	-22.17
PF13_0207	-22.02
MAL6P1.275	-19.71
PFB0585w	-18.76
PFE0145w	-18.45
MAL13P1.319	-17.88
PFL0175c	-17.47
PF14_0697	-17.43
PFI0375w	-15.73
rpoB	-15.66
PF11_0044	-15.52
PF14_0658	-13.43
PFI1585c	-13.27
MAL6P1.203	-12.65
PFL1775c	-12.62
PFL1140w	-12.16
MAL6P1.291	-10.75
PF11_0212	-10.44
PF14_0265	-10.25
MAL13P1.32	-9.85
PFL1120c	-9.59
MAL13P1.42	-9.36
PFL2230c	-9.17
PFE0660c	-9.01
PFB0270w	-8.82
MAL6P1.97	-8.61
PFC0725c	-8.46
PF14_0114	-7.90
PFD0980w	-7.73
MAL6P1.242	-7.49
PFE0635c	-7.16
PFL1350w	-7.09
PF14_0564	-7.09
PFI0330c	-7.07
MAL6P1.285	-7.05
MAL8P1.110	-7.01
MAL6P1.175	-6.73
PFI0920c	-6.04
PF14_0481	-5.92
PFL0305c	-5.85
PF11_0092	-5.75
PF13_0155	-5.61
MAL8P1.27	-4.94
PFI1645c	-4.87
PFE0665c	-4.73
PF11_0172	-4.72
MAL13P1.304	-4.71
PFL1260w	-4.58
PF10_0300	-4.50
MAL8P1.101	-4.47
Clp	-4.32
MAL7P1.20	-4.32
PFC0980c	-4.21
PF11_0229	-4.03
PFL0620c	-3.89
MAL13P1.214	-3.85
PFD0555c	-3.66
PF13_0175	-3.58
MAL6P1.148	-3.58
MAL13P1.329	-3.56
PFI0735c	-3.30
PFI1160w	-3.23
PFA0340w	-3.22
MAL6P1.95	-3.18
PF07_0113	-3.17
PFE1455w	-3.17
PFE0560c	-3.03
PF10_0268	-3.02
PFE1275c	-2.97
PF14_0387	-2.95
MAL13P1.111	-2.88
PF11_0307	-2.84
PFE0050w	-2.77
MAL6P1.205	-2.77
MAL13P1.40	-2.53
PF07_0064	-2.53
PFL1115w	-2.44
PFE0150c	-2.41
MAL8P1.103	-2.38
MAL8P1.13	-2.36
MAL13P1.31	-2.30
PF13_0332	-2.28
PFB0390w	-2.24
PFD0350w	-2.23
PFB0855c	-2.21
PFE0630c	-2.19
PFC0565w	-2.17
PF10_0313	-2.17
MAL7P1.29	-2.17
MAL6P1.217	-2.13
PFI0605c	-2.09
PFL1795c	-2.05
PF10_0040	-1.94
MAL13P1.138	-1.91
MAL13P1.260	-1.89
PF11_0403	-1.87
PF07_0018	-1.80
PFE1040c	-1.80
PFI0720w	-1.72
PFE0320w	-1.71
PF14_0662	-1.67
MAL8P1.141	-1.66
PFL0205w	-1.65
MAL6P1.138	-1.64
PF11_0059	-1.63
PF13_0172	-1.61
PF13_0210	-1.61
PF11_0190	-1.61

Table 5-2 The top 20 drug target genes in the *P. falciparum* proteome

PlasmoDB ID	Drug Target Score (S)	Gene Annotation
PF14_0334	-384.34	NAD(P)H-dependent glutamate synthase, putative
MAL6P1.110	-281.42	transketolase, putative
PF14_0246	-246.58	phosphoenolpyruvate carboxylase, putative
PF14_0641	-190.23	1-deoxy-D-xylulose 5-phosphate reductoisomerase
PFA0225w	-143.36	LytB protein
PF14_0541	-128.15	V-type H(+)-translocating pyrophosphatase, putative
ORF470	-124.01	encodes a well-conserved protein recorded from the plastids of three red algae, O. sinensis and Mycobacteri
PF10_0123	-122.64	GMP synthetase
PFI0355c	-119.39	ATP-dependent heat shock protein, putative
PF08_0063	-112.91	hypothetical protein
PFB0505c	-110.78	beta-ketoacyl-acyl carrier protein synthase III precursor, putative
PF10_0221	-108.04	GcpE protein
PF13_0234	-102.20	phosphoenolpyruvate carboxykinase
PFB0180w	-100.75	5'-3' exonuclease, N-terminal resolvase-like domain, putative
PF14_0133	-100.43	ATP-dependent transporter, putative
PFI1340w	-97.00	fumarate hydratase, putative
MAL6P1.199	-87.82	chorismate synthase
PFI0380c	-75.78	formylmethionine deformylase, putative
PF07_0062	-73.65	GTP-binding translation elongation factor tu family protein, putative
PF11_0337	-72.66	50S ribosomal protein L2, putative

Table 5-S1 Genome collection

Species	Number of annotated ORFs	
Aeropyrum pernix	1841	Archaea
Archaeoglobus fulgidus DSM 4304	2420	Archaea
Halobacterium sp. NRC-1	2622	Archaea
Methanocaldococcus jannaschii	1785	Archaea
Methanococcus maripaludis S2	1722	Archaea
Methanopyrus kandleri AV19	1687	Archaea
Methanosarcina acetivorans C2A	4540	Archaea
Methanosarcina mazei Goe1	3371	Archaea
Methanothermobacter thermautotrophicus str. Delta H	1873	Archaea
Nanoarchaeum equitans Kin4-M	536	Archaea
Pyrobaculum aerophilum str. IM2	2605	Archaea
Pyrococcus abyssi	1896	Archaea
Pyrococcus furiosus DSM 3638	2125	Archaea
Pyrococcus horikoshii	1956	Archaea
Sulfolobus solfataricus	2977	Archaea
Sulfolobus tokodaii	2826	Archaea
Thermoplasma acidophilum	1482	Archaea
Thermoplasma volcanium	1499	Archaea
Acidithiobacillus ferrooxidans ATCC 23270	3172	Bacteria
Agrobacterium tumefaciens str. C58	5402	Bacteria
Aquifex aeolicus VF5	1560	Bacteria
Bacillus anthracis str. Ames	5311	Bacteria
Bacillus cereus ATCC 10987	5844	Bacteria
Bacillus cereus ATCC 14579	5255	Bacteria
Bacillus halodurans	4066	Bacteria
Bacillus subtilis subsp. subtilis str. 168	4112	Bacteria
Bacteroides thetaiotaomicron VPI-5482	4778	Bacteria
Bdellovibrio bacteriovorus HD100	3587	Bacteria
Bifidobacterium longum NCC2705	1729	Bacteria
Bordetella bronchiseptica	4994	Bacteria
Bordetella parapertussis	4185	Bacteria
Bordetella pertussis	3436	Bacteria
Borrelia burgdorferi B31	1640	Bacteria
Bradyrhizobium japonicum USDA 110	8317	Bacteria
Brucella melitensis 16M	3198	Bacteria
Brucella ovis	3382	Bacteria
Brucella suis 1330	3264	Bacteria
Buchnera aphidicola str. APS (Acyrthosiphon pisum)	574	Bacteria
Buchnera aphidicola str. Bp (Baizongia pistaciae)	504	Bacteria
Buchnera aphidicola str. Sg (Schizaphis graminum)	546	Bacteria
Burkholderia mallei ATCC 23344	4888	Bacteria
Campylobacter jejuni RM1221	1841	Bacteria
Campylobacter jejuni subsp. jejuni NCTC 11168	1634	Bacteria
Candidatus Blochmannia floridanus	583	Bacteria
Carboxydothermus hydrogenoformans Z-2901	2645	Bacteria
Caulobacter crescentus CB15	3737	Bacteria
Chlamydia muridarum	911	Bacteria
Chlamydia trachomatis	895	Bacteria
Chlamydophila caviae GPIC	1005	Bacteria
Chlamydophila pneumoniae AR39	1112	Bacteria
Chlamydophila pneumoniae CWL029	1054	Bacteria
Chlamydophila pneumoniae J138	1069	Bacteria
Chlamydophila pneumoniae TW-183	1113	Bacteria
Chlorobium tepidum TLS	2252	Bacteria
Chromobacterium violaceum ATCC 12472	4407	Bacteria
Clostridium acetobutylicum	3848	Bacteria

Clostridium perfringens ATCC 13124	3040	Bacteria
Clostridium perfringens str. 13	2723	Bacteria
Clostridium tetani E88	2373	Bacteria
Colwellia psychrerythraea 34H	4921	Bacteria
Corynebacterium diphtheriae	2291	Bacteria
Corynebacterium efficiens YS-314	2950	Bacteria
Corynebacterium glutamicum ATCC 13032	2993	Bacteria
Coxiella burnetii RSA 493	2045	Bacteria
Dehalococcoides ethenogenes 195	1581	Bacteria
Deinococcus radiodurans	3182	Bacteria
Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough	3531	Bacteria
Enterococcus faecalis V583	3265	Bacteria
Escherichia coli CFT073	5379	Bacteria
Escherichia coli K12	4311	Bacteria
Escherichia coli O157:H7	5341	Bacteria
Escherichia coli O157:H7 EDL933	5324	Bacteria
Fibrobacter succinogenes S85	3275	Bacteria
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	2067	Bacteria
Geobacter sulfurreducens PCA	3445	Bacteria
Gloeobacter violaceus	4430	Bacteria
Haemophilus ducreyi 35000HP	1717	Bacteria
Haemophilus influenzae Rd	1657	Bacteria
Helicobacter hepaticus ATCC 51449	1875	Bacteria
Helicobacter pylori 26695	1576	Bacteria
Helicobacter pylori J99	1491	Bacteria
Lactobacillus johnsonii NCC 533	1821	Bacteria
Lactobacillus plantarum WCFS1	3009	Bacteria
Lactococcus lactis subsp. lactis	2358	Bacteria
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130	3660	Bacteria
Leptospira interrogans serovar lai str. 56601	4727	Bacteria
Listeria innocua	3043	Bacteria
Listeria monocytogenes 4b H7858	3020	Bacteria
Listeria monocytogenes EGD-e	2846	Bacteria
Listeria monocytogenes str. 4b F2365	2821	Bacteria
Mesorhizobium loti	7274	Bacteria
Methylococcus capsulatus Bath	2962	Bacteria
Mycobacterium avium subsp. paratuberculosis str. k10	4350	Bacteria
Mycobacterium bovis subsp. bovis AF2122/97	3920	Bacteria
Mycobacterium leprae	1605	Bacteria
Mycobacterium tuberculosis CDC1551	4187	Bacteria
Mycobacterium tuberculosis H37Rv	3927	Bacteria
Mycoplasma arthritidis 158L3-1	710	Bacteria
Mycoplasma gallisepticum R	726	Bacteria
Mycoplasma mycoides subsp. mycoides SC str. PG1	1016	Bacteria
Mycoplasma penetrans	1037	Bacteria
Mycoplasma pneumoniae	689	Bacteria
Mycoplasma pulmonis	782	Bacteria
Myxococcus xanthus DK 1622	6291	Bacteria
Neisseria meningitidis MC58	2079	Bacteria
Neisseria meningitidis Z2491	2065	Bacteria
Nitrosomonas europaea ATCC 19718	2461	Bacteria
Nostoc sp. PCC 7120	6129	Bacteria
Oceanobacillus iheyensis HTE831	3500	Bacteria
Onion yellows phytoplasma	754	Bacteria
Parachlamydia sp. UWE25	2031	Bacteria
Pasteurella multocida	2032	Bacteria
Photorhabdus luminescens subsp. laumondii TTO1	4683	Bacteria
Pirellula sp.	7325	Bacteria
Porphyromonas gingivalis W83	1909	Bacteria
Prochlorococcus marinus str. MIT 9313	2265	Bacteria
Prochlorococcus marinus subsp. marinus str. CCMP1375	1882	Bacteria
Prochlorococcus marinus subsp. pastoris str. CCMP1986	1712	Bacteria

Propionibacterium acnes KPA171202	2297	Bacteria
Pseudomonas aeruginosa PAO1	5567	Bacteria
Pseudomonas putida KT2440	5350	Bacteria
Pseudomonas syringae pv. tomato str. DC3000	5608	Bacteria
Ralstonia solanacearum	5131	Bacteria
Rhodopseudomonas palustris CGA009	4820	Bacteria
Rickettsia conorii	1374	Bacteria
Rickettsia prowazekii	835	Bacteria
Salmonella enterica subsp. enterica serovar Typhi	4758	Bacteria
Salmonella enterica subsp. enterica serovar Typhi Ty2	4318	Bacteria
Salmonella typhimurium LT2	4527	Bacteria
Shewanella oneidensis MR-1	4471	Bacteria
Shigella flexneri 2a str. 2457T	4068	Bacteria
Shigella flexneri 2a str. 301	4180	Bacteria
Silicibacter pomeroyi DSS-3	4284	Bacteria
Sinorhizobium meliloti	6213	Bacteria
Staphylococcus aureus COL	2678	Bacteria
Staphylococcus aureus subsp. aureus Mu50	2748	Bacteria
Staphylococcus aureus subsp. aureus MW2	2632	Bacteria
Staphylococcus aureus subsp. aureus N315	2624	Bacteria
Staphylococcus epidermidis ATCC 12228	2485	Bacteria
Staphylococcus epidermidis RP62A	2526	Bacteria
Streptococcus agalactiae 2603V/R	2124	Bacteria
Streptococcus agalactiae A909	1966	Bacteria
Streptococcus agalactiae NEM316	2094	Bacteria
Streptococcus mutans UA159	1960	Bacteria
Streptococcus pneumoniae R6	2043	Bacteria
Streptococcus pneumoniae TIGR4	2094	Bacteria
Streptococcus pyogenes M1 GAS	1697	Bacteria
Streptococcus pyogenes MGAS315	1865	Bacteria
Streptococcus pyogenes MGAS8232	1845	Bacteria
Streptococcus pyogenes SSI-1	1861	Bacteria
Streptomyces avermitilis MA-4680	7671	Bacteria
Streptomyces coelicolor A3(2)	8154	Bacteria
Synechococcus sp. WH 8102	2517	Bacteria
Synechocystis sp. PCC 6803	3567	Bacteria
Thermoanaerobacter tengcongensis	2588	Bacteria
Thermosynechococcus elongatus BP-1	2475	Bacteria
Thermotoga maritima	1858	Bacteria
Thermus thermophilus HB27	2210	Bacteria
Treponema denticola ATCC 35405	2767	Bacteria
Treponema pallidum	1036	Bacteria
Tropheryma whipplei str. Twist	808	Bacteria
Tropheryma whipplei TW08/27	783	Bacteria
Ureaplasma urealyticum	614	Bacteria
Vibrio cholerae	3835	Bacteria
Vibrio parahaemolyticus RIMD 2210633	4832	Bacteria
Vibrio vulnificus CMCP6	4514	Bacteria
Vibrio vulnificus YJ016	5024	Bacteria
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	611	Bacteria
Wolbachia endosymbiont of Drosophila melanogaster	1195	Bacteria
Wolinella succinogenes	2044	Bacteria
Xanthomonas axonopodis pv. citri str. 306	4427	Bacteria
Xanthomonas campestris pv. campestris str. ATCC 33913	4181	Bacteria
Xylella fastidiosa 9a5c	2832	Bacteria
Xylella fastidiosa Temecula1	2036	Bacteria
Yersinia pestis biovar Medievalis str. 91001	4142	Bacteria
Yersinia pestis CO92	4067	Bacteria
Yersinia pestis KIM	4449	Bacteria
Anopheles gambiae str. PEST	15212	Eukaryota
Arabidopsis thaliana	26552	Eukaryota
Caenorhabditis elegans	22228	Eukaryota

Candida albicans SC5314	6381	Eukaryota
Drosophila melanogaster	13372	Eukaryota
Encephalitozoon cuniculi	1996	Eukaryota
Eremothecium gossypii	4714	Eukaryota
Giardia lamblia ATCC 50803	9748	Eukaryota
Guillardia theta	632	Eukaryota
Homo sapiens	23546	Eukaryota
Plasmodium falciparum 3D7	5366	Eukaryota
Plasmodium yoelii yoelii	7861	Eukaryota
Rattus norvegicus	20068	Eukaryota
Saccharomyces bayanus 623-6C	4966	Eukaryota
Saccharomyces bayanus MCYC 623	9424	Eukaryota
Saccharomyces castellii	4677	Eukaryota
Saccharomyces cerevisiae	6703	Eukaryota
Saccharomyces kluyveri NRRL Y-12651	2968	Eukaryota
Saccharomyces kudriavzevii IFO 1802	3768	Eukaryota
Saccharomyces mikatae	9057	Eukaryota
Saccharomyces mikatae IFO 1815	3100	Eukaryota
Saccharomyces paradoxus NRRL Y-17217	8955	Eukaryota
Schizosaccharomyces pombe	5008	Eukaryota
Toxoplasma gondii	20904	Eukaryota
Trypanosoma brucei	9154	Eukaryota
Trypanosoma cruzi	25041	Eukaryota

Table 5-S2 Drug target score (*S*) for the complete *P. falciparum* proteome

Due to the large size of this table, the data can be downloaded electronically at the following location: http://derisilab.ucsf.edu/thesisdata/zhu/chapter5_tableS2.pdf .

Table 5-S3 Drug target score (*S*) for known target genes (a positive set)

	EC	PlasmoDB ID	Drug target score (*S*)	Gene annotation	Inhibitor	Pathway
1	1.1.1.267	PF14_0641	-190.23	1-deoxy-D-xylulose 5-phosphate reductoisomerase	fosmidomycin	Isoprenoids biosynthesis
2	2.7.6.3/ 2.5.1.15	PF08_0095	-36.82	dihydropteroate synthase	sulfone/ sulfonamide drugs	Floate biosynthesis
3	1.5.1.3	PFD0830w	123.30	dihydrofolate reductase	pyrimethamine, cycloguanil	Floate biosynthesis
4	3.5.1.88	PFI0380c	-75.78	formylmethionine deformylase, putative	actinonin	Peptide deformylation
5	4.2.3.5	MAL6P1.199	-87.82	chorismate synthase	glyphosate inhibit pathway	Shikamate biosynthesis
6	2.2.1.7	PF13_0207	-22.02	1-deoxy-D-xylulose-5-phosphate synthase		Isoprenoids biosynthesis
7	4.6.1.12	PFB0420w	-69.04	2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase		Isoprenoids biosynthesis
8	1.17.1.2	PFA0225w	-143.36	LytB		Isoprenoids biosynthesis
9	1.17.4.3	PF10_0221	-108.04	GcpE		Isoprenoids biosynthesis
10	1.3.1.9	MAL6P1.275	-19.71	enoyl-ACP-reductase (FabI)	triclosan	Fatty acid biosynthesis II
11	2.3.1.41	PFB0505c	-110.78	beta-ketoacyl-acyl carrier protein synthase (FabH)	thiolactomycin	Fatty acid biosynthesis II
12	2.3.1.41	MAL6P1.165	40.16	3-oxoacyl-(acyl-carrier-protein) synthase i/ii (FabB/F)	thiolactomycin	Fatty acid biosynthesis II

Table 5-S4 Seventy-seven co-ligands of the drug target genes, identified by comparative protein structure modeling or enzymatic annotations

Ligand code	Compound name	Minimum drug target score	Maximum drug target score
2OG	2-oxo-glutaric acid	-384.35	-384.35
ONL	5-oxo-l-norleucine	-384.35	-384.35
D-SEDOHEPTULOSE-7-P	sedoheptulose-7-phosphate	-281.42	-281.42
ERYTHROSE-4P	erythrose-4-phosphate	-281.42	-281.42
DCO	3,3-dichloro-2-phosphonomethyl-acrylic acid	-246.58	-246.58
ADJ	nicotinamide-adenine-dinucleotide-adenylateintermediate	-122.64	-122.64
N-ACETYLNEURAMINATE	n-acetylneuraminate	-110.78	-110.78
1-HYDROXY-2-METHYL-2-BUTENYL-4-DIPHOSPHA	1-hydroxy-2-methyl-2-butenyl 4-diphosphate	-143.36	-108.04
SRM	Siroheme	-108.04	-108.04
GTO	phosphomethylphosphonic acid-guanylate ester	-102.20	-102.20
BB2	Actinonin	-75.78	-75.78
2C-METH-D-ERYTHRITOL-CYCLODIPHOSPHATE	2-c-methyl-d-erythritol-2,4-cyclodiphosphate	-108.04	-69.04
CDF	cytidine-5'-diphosphate	-69.04	-69.04
CHORISMATE	Chorismate	-87.82	-68.50
DAC	2-decenoyl n-acetyl cysteamine	-58.07	-58.07
XYL	d-xylitol	-51.95	-51.95
U	uridine-5'-monophosphate	-51.60	-51.60
THZ	4-methyl-5-(beta-hydroxyethyl)thiazole	-49.75	-49.75
TZE	2-(4-methyl-thiazol-5-yl)-ethanol	-49.75	-49.75
CADAVERINE	Cadaverine	-44.18	-44.18
ACQ	diphosphomethylphosphonic acid adenylate ester	-36.89	-36.89
TMF	5,10-methylene-6-hydrofolic acid	-36.89	-36.89
P-AMINO-BENZOATE	*p*-aminobenzoate	-68.50	-36.82
7,8-dihydropteroate	7,8-dihydropteroate	-36.89	-36.82
DIHYDROPTERIN-CH2OH-PP	2-amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine diphosphate	-36.82	-36.82
AMINO-OH-HYDROXYMETHYL-DIHYDROPTERIDINE	2-amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine	-36.82	-36.82
PMM	pterin-6-yl-methyl-monophosphate	-36.82	-36.82
SAN	Sulfanilamide	-36.82	-36.82
PROPIONATE	propanoate	-31.35	-31.35
AMINO-HYDROXYMETHYL-METHYL-PYR-P	4-amino-5-hydroxymethyl-2-methylpyrimidine-phosphate	-28.88	-28.88
2-D-THREO-HYDROXY-3-CARBOXY-ISOCAPROATE	2-d-threo-hydroxy-3-carboxy-isocaproate	-23.48	-23.48
DEOXYXYLULOSE-5P	1-deoxy-d-xylulose 5-phosphate	-190.23	-22.02

137	1-(o-carboxy-phenylamino)-1-deoxy-d-ribulose-5-phosphate	-143.36	-17.88
CARBOXYPHENYLAMINO-DEOXYRIBULOSE-P	1-(o-carboxyphenylamino)-1'-deoxyribulose-5'-phosphate	-17.88	-17.88
INDOLE-3-GLYCEROL-P	indole-3-glycerol-phosphate	-17.88	-17.88
KCX	lysine nz-carboxylic acid	-36.89	-17.43
SCR	sucrose octasulfate	-9.85	-9.85
IMH	1,4-dideoxy-4-aza-1-(s)-(9-deazahypoxanthin-9-yl)-d-ribitol	-9.01	-9.01
DEOXY-RIBOSE-1P	deoxy-ribose-1-phosphate	-9.01	-9.01
DEOXYGUANOSINE	deoxyguanosine	-9.01	-9.01
RIBOSE-1P	ribose-1-phosphate	-9.01	-9.01
DEOXYINOSINE	deoxyinosine	-9.01	-9.01
A3P	adenosine-3'-5'-diphosphate	-7.73	-7.73
PAP	adenosine-3'-5'-diphosphate	-7.73	-7.73
THZ-P	4-methyl-5-(beta-hydroxyethyl)thiazole phosphate	-49.75	-7.05
AMINO-HYDROXYMETHYL-METHYLPYRIMIDINE-PP	4-amino-5-hydroxymethyl-2-methylpyrimidine-pyrophosphate	-28.88	-7.05
THIAMINE-P	thiamine-phosphate	-7.05	-7.05
16A	cetyl-trimethyl-ammonium	-3.85	-3.85
CPD-406	n-methylethanolamine phosphate	-3.85	-3.85
BIO	biopterin	-3.58	-3.58
P3I	tripolyphosphate	-3.58	-3.58
2-C-METHYL-D-ERYTHRITOL-4-PHOSPHATE	2-c-methyl-d-erythritol-4-phosphate	-190.23	-3.22
CPD-78	4-diphosphocytidyl-2-c-methylerythritol 2-phosphate	-69.04	-2.41
CPD-77	4-diphosphocytidyl-2-c-methylerythritol	-3.22	-2.41
CDM	4-diphosphocytidyl-2-c-methyl-d-erythritol	-2.41	-2.41
HMO	4'-hydroxy-7-methoxyisoflavone	-2.23	-2.23
3-ENOLPYRUVYL-SHIKIMATE-5P	5-enolpyruvyl-shikimate-3-phosphate	-87.82	-1.16
L-1-GLYCERO-PHOSPHORYLCHOLINE	l-1-glycero-3-phosphocholine	-1.60	-0.77
OROTIDINE-5-PHOSPHATE	orotidine-5'-phosphate	-2.19	20.50
DAT	2'-deoxyadenosine-5'-diphosphate	-119.39	95.66
SAI	s-adenosyl-l-homoselenocysteine	-3.85	100.93
DI-H-OROTATE	4,5-dihydroorotate	-17.43	108.78
PROPIONYL-COA	propionyl-coenzyme A	-31.35	155.83
PRPP	5-phospho-α-d-ribose 1-diphosphate	-2.19	157.87
THD	({alpha,beta}-dihydroxyethyl)-thiamin diphosphate 2-[3-[(4-amino-2-methyl-5-pyrimidinyl)methyl]-2-(1,2-dihydroxyethyl)-4-methyl-1,3-thiazol-3-ium-5-yl]ethyl trihydrogen diphosphate	-281.42	159.94
CARBAMYUL-L-ASPARTATE	carbamoyl-l-aspartate	-17.43	162.32
THF	5-hydroxymethylene-6-hydrofolic acid	-34.48	187.68
CIB	2-acetylamino-4-methyl-pentanoic acid [1-(1-formyl-pentylcarbamoyl)-3-methyl-butyl]-amide	-22.17	198.81

HCC	2',4,4'-trihydroxychalcone	-3.85	251.26
XANTHOSINE-5-PHOSPHATE	xanthosine-5-phosphate	-122.64	260.98
MOA	mycophenolic acid	-51.60	260.98
TPS	thiamin phosphate	-51.60	260.98
MAU	n-methyl kirromycin	-73.65	320.63
TPP	thiamine diphosphate	-281.42	342.35
SO1	Sordarin	-73.65	362.38
OCS	cysteinesulfonic acid	-122.64	404.95
ANP	phosphoaminophosphonic acid-adenylate ester	-119.39	1466.70

Table 5-S5 Selected compounds for experimental validation

Due to the large size of this table, the data can be downloaded electronically at the following location: http://derisilab.ucsf.edu/thesisdata/zhu/chapter5_tableS5.pdf .

Table 5-S6 The ideal negative control II (anti-probe set) for the evaluation of toxicity to mammalian cells

Due to the large size of this table, the data can be downloaded electronically at the following location: http://derisilab.ucsf.edu/thesisdata/zhu/chapter5_tableS6.pdf .