Integration and mining of malaria molecular, functional and pharmacological data: how far are we from a chemogenomic knowledge space?

Lyn-Marie Birkholtz, Olivier Bastien, Gordon Wells, Delphine Grando, Fourie Joubert, Vinod Kasam, Marc Zimmermann, Philippe Ortet, Nicolas Jacq, Nadia Saïdani, et al.

Malaria Journal

BioMed Central

Review

Open Access

Integration and mining of malaria molecular, functional and pharmacological data: how far are we from a chemogenomic knowledge space?

Lyn-Marie Birkholtz[1], Olivier Bastien[2], Gordon Wells[3], Delphine Grando[4], Fourie Joubert[3], Vinod Kasam[5], Marc Zimmermann[6], Philippe Ortet[7], Nicolas Jacq[5], Nadia Saïdani[4,8], Sylvaine Roy[9], Martin Hofmann-Apitius[6], Vincent Breton[5], Abraham I Louw*[1] and Eric Maréchal*[4]

Address: [1]Department of Biochemistry and African Centre for Gene Technologies, Faculty of Natural and Agricultural Sciences, University of Pretoria, 0002, Pretoria, South Africa, [2]UMR 5163 CNRS-Université Joseph Fourier, Laboratoire Adaptation et Pathogénie des Microorganismes, Institut Jean Roget, 38700, La Tronche, France, [3]Bioinformatics and Computational Biology Unit, Faculty of Natural and Agricultural Sciences, University of Pretoria, 0002, Pretoria, South Africa, [4]UMR 5168 CNRS-CEA-INRA-Université Joseph Fourier, Département Réponse et Dynamique Cellulaires; CEA Grenoble, 17 rue des Martyrs, 38054, Grenoble Cedex 09, France, [5]Laboratoire de Physique Corpusculaire de Clermont-Ferrand, CNRS-IN2P3, Campus des Cézeaux, 63177 Aubière Cedex, France, [6]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, 53754 Sankt Augustin, Germany, [7]Département d'Ecophysiologie Végétale et de Microbiologie; CEA Cadarache, 13108 Saint Paul-lez-Durance, France, [8]UMR 5539 CNRS-Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier cedex 05, France and [9]Laboratoire de Biologie, Informatique et Mathématiques; Département Réponse et Dynamique Cellulaires; CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France

Email: Lyn-Marie Birkholtz - lynmarie.birkholtz@up.ac.za; Olivier Bastien - ol.bastien@wanadoo.fr; Gordon Wells - gordon.wells@gmail.com; Delphine Grando - delphine.grando@cea.fr; Fourie Joubert - fourie.joubert@bioagric.up.ac.za; Vinod Kasam - kasam@clermont.in2p3.fr; Marc Zimmermann - marc.zimmermann@scai.fraunhofer.de; Philippe Ortet - portet@cea.fr; Nicolas Jacq - jacq@clermont.in2p3.fr; Nadia Saïdani - nsaidani@univ-montp2.fr; Sylvaine Roy - sylvaine.roy@cea.fr; Martin Hofmann-Apitius - martin.hofmann-apitius@scai.fhg.de; Vincent Breton - breton@clermont.in2p3.fr; Abraham I Louw* - braam.louw@bioagric.up.ac.za; Eric Maréchal* - eric.marechal@cea.fr

* Corresponding authors

Published: 17 November 2006

Malaria Journal 2006, **5**:110 doi:10.1186/1475-2875-5-110

Received: 29 September 2006
Accepted: 17 November 2006

This article is available from: http://www.malariajournal.com/content/5/1/110

Abstract

The organization and mining of malaria genomic and post-genomic data is important to significantly increase the knowledge of the biology of its causative agents, and is motivated, on a longer term, by the necessity to predict and characterize new biological targets and new drugs. Biological targets are sought in a biological space designed from the genomic data from *Plasmodium falciparum*, but using also the millions of genomic data from other species. Drug candidates are sought in a chemical space containing the millions of small molecules stored in public and private chemolibraries. Data management should, therefore, be as reliable and versatile as possible. In this context, five aspects of the organization and mining of malaria genomic and post-genomic data were examined: 1) the comparison of protein sequences including compositionally atypical malaria sequences, 2) the high throughput reconstruction of molecular phylogenies, 3) the representation of biological processes, particularly metabolic pathways, 4) the versatile methods to integrate genomic data, biological representations and functional profiling obtained from X-omic experiments after drug treatments and 5) the determination and prediction of protein structures and their molecular docking with drug candidate structures. Recent progress towards a grid-enabled chemogenomic knowledge space is discussed.

Background

Malaria is a life-threatening disease affecting half a billion humans in underdeveloped and developing countries. Its global heartland is Africa, with an appalling death toll of 1 to 2 million people every year [1]. Endemic malaria ranges from a permanent incidence in sub-Saharan and equatorial Africa, to a seasonal but recently escalating prevalence in Southern Africa [2]. Four species of malaria parasites can infect humans *via* mosquito transmission: *Plasmodium falciparum* (the species that causes the greatest incidence of illness and death) as well as *Plasmodium vivax*, *Plasmodium ovale*, and *Plasmodium malariae*. They belong to the Apicomplexa phylum, which contains many other parasitic protists of medical and veterinary importance [3].

Malaria was eradicated from temperate regions following concerted preventative sanitary actions and after important insecticide campaigns and systematic treatments with available drugs, *i.e.* quinine and chloroquine [4-6]. The prophylactic programmes of the 1950's and 1960's, essentially based on insecticide and drug treatments, failed to control malaria in subtropical areas [7]. Resistance to chloroquine spread rapidly [8,9]. Subsequent attempts to achieve progress in malaria prophylaxis have been characterized by the failure of vaccine development, withdrawal of some insecticides because of toxicity and negative environmental impact, the alarming spread of mosquito resistance to insecticides and of resistance of *Plasmodium* to the very few drugs that have been developed [9-11]. The promise of an effective vaccine is as distant as ever [12]. Current efforts focus on chemotherapy using artemisinin, an antiplasmodial molecule from *Artemisia annua*, and derivatives which can be produced efficiently and cheaply. However, the scientific community is worried that plans for the extensive use of artemisinin might be ruined by emergence of the parasitic resistance it will almost certainly trigger, sooner or later [13-15]. Given the small number of available drugs and the resistance they have already induced, discovery of new targets and of new drugs remains a key priority.

A major landmark in the history of malaria was the launch of a collaborative genomic sequencing programme in 1996 [16-21]. In November 2002, the complete genome of the 3D7 strain of *P. falciparum* [22] and a whole genome-shotgun of *Plasmodium yoelii yoelii* [23] were released, followed by whole-genome shotguns of *Plasmodium berghei*, *Plasmodium chabaudi*, and *P. vivax*, and the genomic sequencing of *Plasmodium gallinaceum*, *Plasmodium knowlesi* and *Plasmodium reichenowi* still in progress [21,24]. This unprecedented effort to sequence genomes of eukaryotic pathogens was a technical challenge, because the extreme compositional bias of *Plasmodium* DNA (>80% A+T in *P. falciparum*) accounted for the instability of genomic fragments in bacteria [17-19] and complicated assembly of contigs [19]. Among eukaryotes, the *Plasmodium* genus is, therefore, the best documented at the genomic sequence level, with well-established syntenic relations. At the level of the Apicomplexa phylum, additional complete genomes of *Cryptosporidium*, *Theileria* and *Toxoplasma* have been either released or announced [25,26].

All *Plasmodium* molecular data have been collected and organized in the PlasmoDB public database as early as sequencing outputs were made available [27-30]. The architecture of the relational database was designed following biologically relevant relationships, *i.e.* the "gene to mRNA to protein" dogma, using the Gemomics Unified Schema [28,29], and ensures that gene loci are linked to annotation using the Gene Ontology standards [30]. A genome browser allows navigation along chromosome sequences and the viewing of multiple *Plasmodium* species at one glance, based on syntenies. Predictions of protein domains, post-translational modifications, subcellular targeting sequences, etc. are included. Furthermore, PlasmoDB is currently the only site where molecular data are (1) tentatively clustered based on homology, (2) linked to generic schemes designed to view metabolic pathways, and (3) linked to X-omic functional information (transcriptome, proteome, interactome). Any biologist can exploit these integrated data with basic or combined queries [27-30], and this is the first resource designed to help the scientific community to turn genomic data hopefully into gold, *i.e.* appropriate biological knowledge that can accelerate the design and introduction of new therapeutic strategies. PlasmoDB operates inside ApiDB, a master web portal for apicomplexan genomes [29,32].

Contrasting with this integrated and user-friendly access to molecular and functional data, the proportion of genes for which a biological function has been inferred appears like a curse. Only 34 % of the *P. falciparum* genes could be assigned a function, based on detected sequence homology with characterized genes from other organisms [22,33,34]. The most-sequenced eukaryotic group appears, therefore, as the worst functionally annotated. The first version of the *P. falciparum* genome was estimated to code for 5,268 proteins, out of which 3,208 did not have any significant similarity to proteins in other organisms to justify provision of functional assignment. This proportion of uncharacterized genes was further increased by 257 additional sequences, which had significant similarity to proteins, described as hypothetical, in other organisms [22]. The 2005 updated EMBL version of the 3D7 complete genome (generated at Sanger Institute, The Institute for Genomic Research, and Stanford University; version 2.1) was still predicted to encode as many as 3,548 hypothetical proteins (65.6 % of total). This figure

is the worst ever recorded for a eukaryotic genome (Table 1) and a clear limitation to any *in silico* exploration of the malaria biology.

Since absence of evidence is not evidence of absence, the scientific community faces a serious epistemological problem when trying to derive conclusions from a small genome (the number of genes is similar to that of yeast), in which two thirds cannot be used in meaningful analyses. A sustained effort is required to improve functional annotation methods and to contribute to the PlasmoDB gene descriptions. Different teams in the world have attempted to address this problem. This paper provides an overview of some of the theoretical and practical developments that were introduced to improve detection of similarities between *Plasmodium* sequences and distantly related organisms.

A second difficulty, also related to sequence homology detection, is the reconstruction of molecular phylogenies for malarial genes. Accurate molecular phylogenies are particularly important since the *Plasmodium* cellular organization is the result of multiple endosymbioses involving an ancestral alga [3,35,36], at the origin of a plastid relic, *i.e*. the apicoplast [37-43]. Comparative phylogenomics focusing on malarial plant-like genes proved to be a valid strategy to detect potential targets for herbicides that act as antiplasmodial [44]. *In silico* analyses combining molecular phylogeny and targeting sequence prediction allowed a first rationalized mining of the apicoplast function [45]. The identification of all biological processes that have been inherited through lateral gene transfers from the ancestral alga (the algal sub-genome) is one of the most important outputs one expects from comparative phylogenomics. Although molecular phylogenies

Table 1: Comparison of *Plasmodium falciparum*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Homo sapiens* genomic statistics

	Plasmodium falciparum	*Saccharomyces cerevisiae*	*Arabidopsis thaliana*	*Homo sapiens*
Genome general statistics				
No of chromosomes	14	16	5	22 + X/Y
Size (bp)	22,853,764	12,495,682	115,409,949	3,272,187,692
average (A+T) %	80.6	61.7	65.1	59.0
Estimated number of genes	5,268	5,770	25,498	31,778
Average gene length	2,283	1,424	1,310	1,340
% of coding genome	53	66	29	9
Initial annotation based on sequence similarity (BLAST or *Smith-Waterman *E-values*)				
Proportion of predicted protein sequences:				
- having a detectable similarity to sequences, in other organisms, of known function at the initial genome release date.	34 %	75 %	69 %	59 %*
- without any detectable similarity to sequences in other organisms at the initial genome release date, i.e. "no BLASTP match to known proteins" (estimates based on published data and local BLAST searches).	61 %	< 8 %	< 20 %	15 %*
- of totally unknown function (hypothetical proteins = with similarity to sequences of unknown function + without any detectable similarity to sequences in other organisms).	66 %	16 %	31 %	41 %*
Average characteristics of open reading frames				
Exons:				
No per gene	2.39	1.05	5.18	12.1
(A+T) %	76.3	60	55	52
average length	949	1356	253	111
Introns:				
(A+T) %	86.5	64	66	60
Intergenic regions:				
(A+T) %	86.4	64	66	60

Presented data compile information from [22] for *Plasmodium falciparum*, [190] for yeast (completed with statistics made available *via* the Comprehensive Yeast Genome Database website, [191]), the Arabidopsis genome initiative [192] for Arabidopsis, and the International Human Genome Sequencing Consortium [193] and [194] for Human (completed with statistics made available *via* Ensembl, [195]). These statistics at the complete genome release date have been continuously updated since then.

of a few genes can be achieved with conventional methods, combined with expert visual analysis, it is difficult to carry out high-throughput phylogenetic determination at a genomic scale. This paper summarizes how the question of automatic assessment of orthologies has been addressed, particularly with the OrthoMCL [46] and TULIP [34,47] approaches.

A third difficulty is the representation of knowledge of malaria parasite biology. Description of gene function follows the guidelines of the Gene Ontology (GO) structured vocabulary [31]. GO is a standard adopted by all the scientific community to circumvent problems raised by heterogeneous key word annotations. GO also complies with *in silico* management (and mining) of information. Genes coding for enzymes can further be linked to metabolic scheme(s) designed using generic methods, *i.e.* KEGG [48] or MetaCyc [49], or specifically designed for *Plasmodium*, *i.e.* the Malaria Parasite Metabolic Pathways (MPMP) [50,51]. Pathways based on generic methods do not include representations for cell compartmentalization. This information, which has been included in the MPMP, is essential to understand the metabolism of water-soluble intermediates inside and between different cell compartments, and is critical for pathways involving lipophilic compounds localized in disconnected membranes. Eventually, knowledge should be represented in a way that is usable for the organization and analysis of functional data. Here, the compliance of each approach with these theoretical and practical constraints, is discussed.

Fourthly, the difficulty of organizing molecular data into knowledge representations becomes more pronounced in the analysis of global datasets arising from X-omic (transcriptome, proteome, interactome) experiments. The strategies, methods and tools, which have been used or designed in order to link malarial molecular and functional data in the most versatile ways, were examined. In particular the MADIBA tool, developed for local analyses of transcriptomic outputs, is described.

A fifth difficulty, once a target gene has been identified for possible antimalarial intervention, is the process of reaching a decision on the entry into costly drug- or vaccine-development programmes. Besides experimental validation criteria, *in silico* experiments can be a useful tool to help characterize a possible new target. These include determination of protein three-dimensional structure and determination of possible binding ligands through *in silico* protein-ligand docking. To date, the structures of less than 70 malaria proteins have been determined and made available *via* the Protein Data Bank [52]. This review provides arguments in favour of a repository for all known malaria resolved protein structures and structural models, which should be initiated, curated and maintained. No prediction of protein druggability has been investigated. Access to such repository might be invaluable for drug discovery projects: virtual screening of hundreds of thousands of potential drugs, making use of protein structures of a whole family has been achieved using computer grid resources with the WISDOM I project [53], and oncoming WISDOM II.

This review does not pretend to provide any complete panorama on malaria molecular, functional and therapeutical genomics or to introduce panaceas. Important difficulties for global analyses and high throughput approaches were addressed here, as five major challenges for the future, *i.e.*, 1) comparisons of the compositionally atypical malaria gene and protein sequences, 2) high throughput molecular phylogeny assessment, 3) usable and interoperable representations of metabolic and other biological process, 4) versatile and local integration of molecular and functional data obtained from X-omic experiments, and 5) determination and prediction of protein structure and subsequent virtual ligand screening on candidate therapeutical targets. Linking protein structures with ligand structures is a pivotal step for a chemogenomics knowledge base, in which functional and structural knowledge deriving from malaria genomics and post-genomics might be connected with the space of small molecules containing known and potential drugs.

Integration of malaria genomic, post-genomic data and chemical information: current status and future challenges

Sequence comparisons of compositionally-biased and insert-containing malaria genes and proteins. The extreme A+T bias in *Plasmodium* DNA has been a recurrent problem for malarial genomics and post-genomics. It was responsible for instability of genomic segments in *E. coli* and difficult assembly during the sequencing process [19]. It implied debated modifications and combinations of automatic gene detection methods for open reading frame prediction [54,55]. The nucleotide bias has been demonstrated to be responsible for the protein composition bias in *P. falciparum* [56-58]. Some parasites of the *Plasmodium* genus, like *P. vivax*, do not show such a strong compositional bias at the DNA level, but their protein sequences appear to be also compositionally divergent from average.

Combined with the frequent insertions seen in malarial proteins, the amino acid compositional bias is critical for routine sequence comparison methods, particularly because it can compromise the statistical analyses and sorting of BlastP alignments [58]. Indeed, an alignment algorithm comes with a statistical model implemented in the code, particularly in the Blast package [59], on which users rely to assess the significance of the alignment, and to sort them. It is, therefore, difficult to discuss the current

view on sequence comparison methods and statistics independently. Two major statistical models are used to test alignment scores. The most common test is an estimate of the *E-value* (short for Expectation value), *i.e*. the number of alignments one expects to find in the database by chance, with equivalent or better scores. It can be determined from the complete distribution of scores. The BlastP associated statistics defined by Karlin and Altschul [59] are based on the probability of an observed local alignment score according to an extreme value distribution. The validation of the Karlin-Altschul *E-value* computation model requires two restrictive conditions: first, individual residue distributions for the two sequences should not be 'too dissimilar' and second, sequence lengths 'should grow at roughly equal rates' [59]. Validity restrictions listed here are fully acceptable when dealing with protein sequences of average lengths and amino acid distribution, and BlastP, besides its constitutive limits in detecting short similarities, is a good compromise for batch analyses of genomic outputs. However, the compositionally biased proteome of *P. falciparum* fall outside of the validity domain for a BlastP comparison with unbiased sequences [58,60]. One of the reasons explaining that 60 % of the *Plasmodium* sequences did not have any apparent homology with sequences from other genomes may not be that most malarial genes are unique in the living world, but that the BlastP semi-automatic annotation procedure was technically limited. Some missed sequences could be retrieved by adding protein structural information (such as hydrophobic cluster analysis, [61]), but these methods require visual expertise and cannot easily be automated. Iterating the BlastP procedure has also proven to be helpful in detecting missed homologies [62], providing evidence for the initial failure of alignment significance detection.

An alternative method to assess the relevance of a pairwise alignment was introduced by Lipman and Pearson [63]. It uses Monte Carlo techniques to investigate the significance of a given score calculated from the alignment of two real sequences. It can be used to sort results obtained by any comparison methods, including BlastP, although this has not yet been achieved at a massive scale. It is currently used to estimate the probabilities of Smith-Waterman comparisons [64]. The asymptotic law of *Z-value* was shown to be independent of sequence length and amino acid distribution [65] and is fully valid for malaria sequence comparisons. Bastien et al. [60] demonstrated the TULIP theorem (theorem of the upper limit of a score probability) assessing that *Z-values* can be used as a statistical test and a single-linkage clustering criterion. In practice, a *Z-value* table can be analysed using the TULIP theorem to detect pairs of proteins that are probable homologues following a *Z-value* confidence cutoff. For instance, a *Z-value* above 10 allows an estimate that the alignment is significant with a statistical risk of $1/Z\text{-value}^2$, *i.e*. 0.01. A version of the BlastP algorithm, implemented with *Z-value* statistics, should be helpful to refine malaria sequence comparisons.

Additional improvement of automatic annotation procedure are expected, in particular by combining sequence comparisons with GO term associations (*e.g*. GOtcha; [33]) a complementary approach to the annotation based on the combination of GO terms with functional X-omic response patterns (*e.g*. Ontology-based pattern identification – OPI – following the guilt-by-association principle; [66]). In the last section of this article, improvements with information obtained from multiple alignments are also presented (see below).

Genome-scale assessment of malaria molecular phylogenies. The amino acid compositional bias and high insert content of malaria proteins is also a disturbing factor when attempting to reconstruct phylogenies. Conventional methods used for phylogeny reconstructions based on multiple alignments can be used in conjunction with visual judgment,- [67], with qualitative decisions on how protein segments "align well". However, such manual pretreatment cannot be undertaken for all known genes. Alternatively, high throughput molecular phylogenies can be derived from massive all-against-all comparisons, based on pairwise alignments [68]. The questions of the statistical accuracy and maintenance of high throughput phylogenetic reconstruction are critical when including compositionally atypical and high insert containing sequences.

The output of an all-by-all comparison of n protein sequences is an $n \times n$ table. According to the output table processing, it can be either totally recomputed at each database update, or stored and updated by computing complementary tables. Information is extracted from the output table to help reduce complexity and diversity at the sequence level. Sets of sequences sharing features are named "clusters" [69].

A first massive comparison project, OrthoMCL, was designed to cluster malaria genes based on their sequence similarity with genes of 55 other genomes (> 600,000 sequences), using the BlastP algorithm and Karlin-Altschul *E-value* statistics to build the all-against-all comparison table [46,70]. As mentioned above, and discussed recently [68] for massive comparisons based on BlastP/*E-values*, *i.e*. COG [71], Tribe [72], ProtoMap [73], ProtNet [74], SIMAP [75] and SYSTERS release 4 [76], there is no theoretical support to justify that an *E-value* table can be converted into a rigorous and stable metric. The handling of the output $n \times n$ table of *E-values* requires pragmatic post-processing normalization, including asymmetric cor-

rections of *E-values* obtained after permutation of the two aligned sequences or consensus *E-value* computation after alignment with different algorithms. The *E-value* table can be converted into a Markov matrix (*e.g.* Tribe, SIMAP, OrthoMCL), or a close graphic equivalent, *i.e.* graphs connecting protein entries with *E-values* as weights for graph edges (*e.g.* COG, ProtoMap, SYSTERS), a representation that has been exploited in the OrthoMCL project. The protein sets are organized either by detecting graphs and subgraphs following pragmatic rules, with granularities depending on *E-value* thresholds, or by distance clustering using *E-value* as a pseudo-metrics, or by Markov-random-field clustering. None of the organization of the protein sequences obtained through these methods can be named a spatial projection, and none of the obtained clusters can be represented as a phylogenetic reconstruction. Eventually, the economy of computing *E-values* in an all-by-all comparison experiment is lost in the updating process that requires a complete re-calculation. In spite of these drawbacks, massive comparisons based on BlastP/*E-values* have been undertaken because they were less CPU-demanding than other methods. The OrthoMCL clustering method based on BlastP/*E-value* represents therefore a pragmatic reduction of the protein diversity, and phylogeny reconstruction require post treatments within each clusters. OrthoMCL flags probable orthologous pairs identified by BlastP as reciprocal best hits across genomes. Access to OrthoMCL groups is linked to the PlasmoDB GUS underlying database, allowing multiple queries with other PlasmoDB data and information, and allowing additional cross-species/cross-phylum profiling of the BlastP/*E-value*-supported orthologues.

A more CPU-demanding alternative method for massive all-against-all protein sequence comparison uses Smith-Waterman/*Z-values* rather than BlastP/*E-values*. This method has been initiated for the ClusTR protein sequence clustering, underlying the UniProt/Integr8 knowledge base at the European Bioinformatics Institute, EBI [77]. Because of the properties of the *Z-value* statistics detailed above, it is the solution of choice when comparing compositionally biased and high-insert containing sequences. Additionally, for any set of homologous proteins, it is possible to measure a table of pair-wise divergence times and build phylogenetic trees using distance methods [47]. These trees are called TULIP trees. TULIP trees were compared to phylogenetic trees built using conventional methods, for instance the popular PHYLIP [78] or PUZZLE [79] methods based of multiple sequence alignments. TULIP trees proved to perform as well in any unbiased sets of proteins. Moreover, some phylogenetic inconsistencies in trees built with multiple-alignment based methods, particularly including subsets of compositionally biased sequences, or with low bootstrapping values, could be spectacularly solved with the TULIP tree [47]. An advantage of the phylogenetic inference from the CSHP over that obtained from multiple alignments lies precisely in the TULIP tree construction from pair-wise alignments. Whereas the addition or removal of a sequence can deeply alter the multiple alignment result, and the deduced phylogeny, the *Z-value* and divergence time tables that serve to reconstruct the TULIP trees are the result of a Monte Carlo simulation, which is a convergent process at the level of the pair-wise comparison and is not altered by database updates. As a result, whereas a phylogenetic database computed from multiple alignments would require a complete and increasing computation for any update, the TULIP tree calculation simply requires the calculation of the {new}-by-{old} and {new}-by-{new} *Z-values* and divergence times. A mapping of each *Plasmodium* sequence can be obtained and updated following all-against-all pairwise comparisons based on *Z-value* statistics. The CPU-cost required by the Smith-Waterman comparison method and by the Monte-Carlo simulation used to compute *Z-values* will be compensated in the future by implementing *Z-values* on the BlastP heuristics. A high-throughput assessment of molecular phylogenies of *Plasmodium* genes based on BlastP/*Z-value*, including all recorded genes in public databases, will therefore be feasible and upgraded at the pace of public database updates.

User access to molecular phylogenies, which has been designed in the OrthoMCL project in a very practical and user-friendly way, is essential to mine the genome for clues to therapeutic opportunities. The most obvious approach is to detect protein sequences that are excluded or diverge strongly from the mammalian proteome. More subtly, the question of the plant/algal sub-genome of Apicomplexans has been demonstrated to be a source of therapeutic targets for herbicidal drugs (*e.g.* apicoplast lipid – fatty acid, isoprenoid -syntheses, plant-like targets localized outside plastids – folate metabolism, tubulin -, etc.). Criteria for confirmation of a plant/algal sub-genome in *P. falciparum* include molecular similarities with plant genes, and will therefore benefit from future progress in high-throughput molecular phylogenies. Information on sequence and sequence similarity are not sufficient to highlight functions that are, for instance, unique to plants: they have to be linked to appropriate knowledge representation of the biological function of each gene and each process in which gene products play their roles.

Knowledge representations of the biological function

Having in hand a table of gene entries, with summarized annotations is not sufficient to handle genomic information. Data can be organized based on biological principles so as to reflect current knowledge at best, and to be viewed at a glance in global X-omic experiments, or for comparative purposes. The question of knowledge representation in reductionist terms (*i.e.* based on the fact that some lev-

els of knowledge can be reconstructed by the integration of parts of knowledge of lower levels) is a difficult epistemological question and can hardly be debated here. Current data integration strategies are dependent on available consensus methods, with their compromises and imperfections that have been slowly adopted by the scientific communities. Biological (and chemical) knowledge is primarily produced in the form of research publications, books, patents and other un-structured texts, and since two decades in structured databases (biologists are used to fill Genbank forms prior to paper submissions). Data and information can be organized as semantic networks, following combinations of ontologic hierarchies and praxeologic schemes [80]. In brief, an ontological hierarchy is designed to organize entities (here biological or chemical entities) following inclusion/subsumption principles ("A" is part of, is a component of, etc. "B"). Best known examples are the taxonomic trees ("species" is part of "genus", etc.) or the Gene Ontology, or GO [31], although this later was built from a lose definition of the ontology (see below). A praxeologic scheme allows a representation of the activity or function (praxis) (for instance enzymatic activities caused by proteins, assembly of molecular structures, biological effects caused by drugs, etc.) based on the transformation or alteration of an entity into another, through time ("C" is converted into "D"). By contrast with ontologies, which should be stable and non-conflicting hierarchies, praxeologies can describe cyclic processes and can vary over time. Best know examples are metabolic graphs and fluxes, and their variations in different physiological conditions. Interestingly, the enzymatic activities can be organized following hierarchical principles (e.g. the EC numbers proposed by the Enzyme Commission): a praxeologic scheme such as a metabolic pathway is therefore linked to one or more ontologies for the metabolites (hierarchical categorization of molecules) and for enzymes (hierarchy following the GO, the EC, etc.). In practice, although they are of very distinct nature, ontologic hierarchies and praxeologic schemes are handled by bioinformaticians as "semantic networks" and relational graphs *senso lato*.

Following recommendations of the Gene Ontology (GO) consortium [31], malaria gene function was defined using a semantic network organized in three hierarchical axes, *i.e.* "molecular function", "biological process" and "cellular component", according to a controlled hierarchical vocabulary. The GO-based annotation circumvents the problems raised by heterogeneous key word annotations, allowing subsequently a cross-species comparison with genomes annotated similarly, and complies with *in silico* management requirements. Genes thus annotated can further be embraced in higher order representations of biological knowledge.

Concerning enzymes, entries can be linked to graphical representations of the metabolic reactions they catalyze. To that end, in the same way genes were defined following the GO procedure, enzyme substrates and products are defined following a chemical ontology (CO), and the reaction itself has to be defined following an enzymatic reaction ontology. A reaction can be viewed as a small graph in which the enzyme is associated with the line that connects the nodes corresponding to the substrates and products. Metabolic pathways are connected by shared nodes, and can be viewed at different scales. Current representations of malaria metabolism have been made available using generic methods, *i.e.* KEGG [48,81] or MetaCyc [49,82,83], or specifically designed for *Plasmodium*, *i.e.* the Malaria Parasite Metabolic Pathways (MPMP) coordinated by Hagai Ginsburg at the Hebrew University of Jerusalem [50,51].

The Kyoto Encyclopedia of Genes and Genomes (KEGG) resource provides a set reference of metabolic schemes, manually designed so as to represent all possible primary metabolic reactions, and formatted with the KEGG Markup Language (KGML) [48]. These reference schemes can be explored on the KEGG web portal, with a very clear view of the global metabolic map, connecting all pathways. Organism-specific schemes are generated, based on sequence similarities (using the KEGG Orthology, or KO orthologue identifier) with references to the KEGG Gene catalogs. Thus 79 metabolic schemes have been generated for *P. falciparum*. Each scheme provides links to metabolite (substrates, products, co-substrates) information, and enzyme descriptions following the Enzyme Commission (EC) classification. EC numbers give access to multiple sequence alignments, protein motifs, genomic mapping, links to Genbank, UniProt, PDB, etc. Tilling KEGG schemes from different organisms highlights metabolic similarities and differences, and could be of help for antimalarial purposes, highlighting for instance metabolic reactions occurring in *Plasmodium* and not in humans. However, the maps are designed based on the enzymatic reactions and, for instance, the fatty acid synthesis due to the type 1 fatty acid synthase (FASI, a multiprotein complex) from the human cell cytosol, strictly overlaps with the type 2 fatty acid synthase (FASII dissociated enzymes) from the *Plasmodium* apicoplast. Information on the protein structure and cell compartmentalization of the enzymes would have been sufficient to distinguish between FASI and FASII. Furthermore, metabolites that are generated in one compartment (*i.e.* diacylglycerol generated in one of the numerous cell membranes) may not be available for an apparent downstream reaction occurring in another compartment. Thus, used without caution, the KEGG schemes may seem to be fully valid for the entire living world, with a misleadingly clear and fully cross-connected global overview of metabolism, and can

lead to unrealistic representations. The missing enzymes, which have been experimentally assayed, are not shown, and it is unclear whether gaps within pathways are due to absent enzymes or to incomplete data. Additionally, the KEGG representation is not intended for the design of schemes other than those pre-defined. KEGG outputs need to be reexamined for accuracy of interpretation (see below).

As an alternative, the design of MetaCyc schemes for *Plasmodium*, called PlasmoCyc, has been initiated [49,82]. As for KEGG, a reference of the complete metabolic pathways has been designed manually and loaded in a MetaCyc database. In contrast with the fully connected overview of the KEGG metabolic map that can make the user overconfident, the MetaCyc global view of metabolism is fragmented, reflecting knowledge gaps, incomplete design of some pathways (such as the tricky pathways for lipid syntheses), and the versatility of the MetaCyc tool for implementation of new schemes. Using *P. falciparum* gene annotation and information of the MetaCyc reference database, the PathoLogic module of the Pathway Tool Software [84] allows the generation of a Pathway/Genome Database (PGDB). A total of 113 metabolic pathways (complete or fragmented) have been generated for *P. falciparum*. As in KEGG, each graph gives access to metabolite, reaction and enzyme information. The Gene-Reaction Schematic (GRS) representation allows a visualization of the relation between the genes, the enzymes, the catalyzed reactions, even in the case of complex or multienzymatic proteins. This model is useful to distinguish for instance FASI and FASII proteins. It is further useful for *Plasmodium* proteins that are often multienzymatic. For each reference pathway, the occurrence or absence of enzyme homologues in the *Plasmodium* genome is documented. As in the case of KEGG, compartmentalization information is missing from the pathway model, in spite of some effort to highlight some specific pathways (for instance the apicoplast fatty acid synthesis).

A synthetic representation of *Plasmodium* biological processes can further be viewed at the Malaria Parasite Metabolic Pathways web portal [50]. Among 120 schemes representing numerous cellular processes, half represent metabolic pathways and were fully designed for malaria researchers. All schemes were built using KEGG pathways, cleaned of irrelevant information and curated by international experts. The quality of each representation is very high and specialized, and benefits from a sustained effort in biological and molecular investigation and validation. Missing data are concisely documented. Most importantly, enzyme subcellular localization is shown. Graphs are not automatically generated, but drawn by experts. While MetaCyc graphs are self generated taking into account all information of the underlying database, the MPMP scheme are not self generated: in particular, the subcellular localization representation is not generated automatically from a subcellular attribute filled in the underlying database. Thus, although the MPMP representations are of higher quality, their update is not dynamic and cannot be used for *in silico* graph-based treatments.

From this short overview, it is clear that metabolic representations should be carefully used. On one hand, generic approaches (KEGG, MetaCyc) do not include cell compartmentalization data and the chemical ontology for lipids is not finished. On the other hand, the curated MPMP representations are based on high-quality data but they are static, with no graph management tools, giving click-access to remote information. Future challenge will be to design underlying models for building graphs from genomic data (following the GO), metabolite data (following a stable Chemical Ontology), reaction connections and compartmentalization information for both gene products and metabolites. Construction of PlasmoCyc benefited from information of the MPMP [50,82], and MPMP was designed after cleaning KEGG schemes. Models for other knowledge bases and graphs should benefit from the important effort of the MPMP in defining metabolic and other biological processes.

Connecting functional schemes and ontologies with post-genomic global functional analyses

The application of functional genomics strategies to assign functionality to each gene product of an organism has recently attained increased attention in the field of post-genomics research of *Plasmodia*, includes the understanding of the transcriptome, proteome and interactome of the parasite to elucidate mode-of-action of inhibitory compounds, allow optimization of such inhibitor activities, explain resistance mechanisms to known drugs, chemically validate potential drug targets and ultimately identify and/or functionally describe new drug targets (Figure 1). Some malaria X-omic studies did not meet the initial expectations, such as transcriptome analyses of *P. falciparum* in which the modification of the transcriptomic pattern to different treatments was particularly low. Some others were technically highly biased, such as interactome analyses in which results from protein fragments are recorded together with results from intact proteins, and raise imperfect, but interesting sets of data. In this part, an overview of malaria X-omic analyses is given, with some insights on the current strategies for *in silico* integration of their functional outputs with genomic knowledge.

The *P. falciparum* transcriptome has been extensively investigated resulting in comprehensive profiles of transcript expression throughout the complete life cycle of the parasite [85,86]. The overall conclusions demonstrated that the majority (~87%) of the predicted genes are

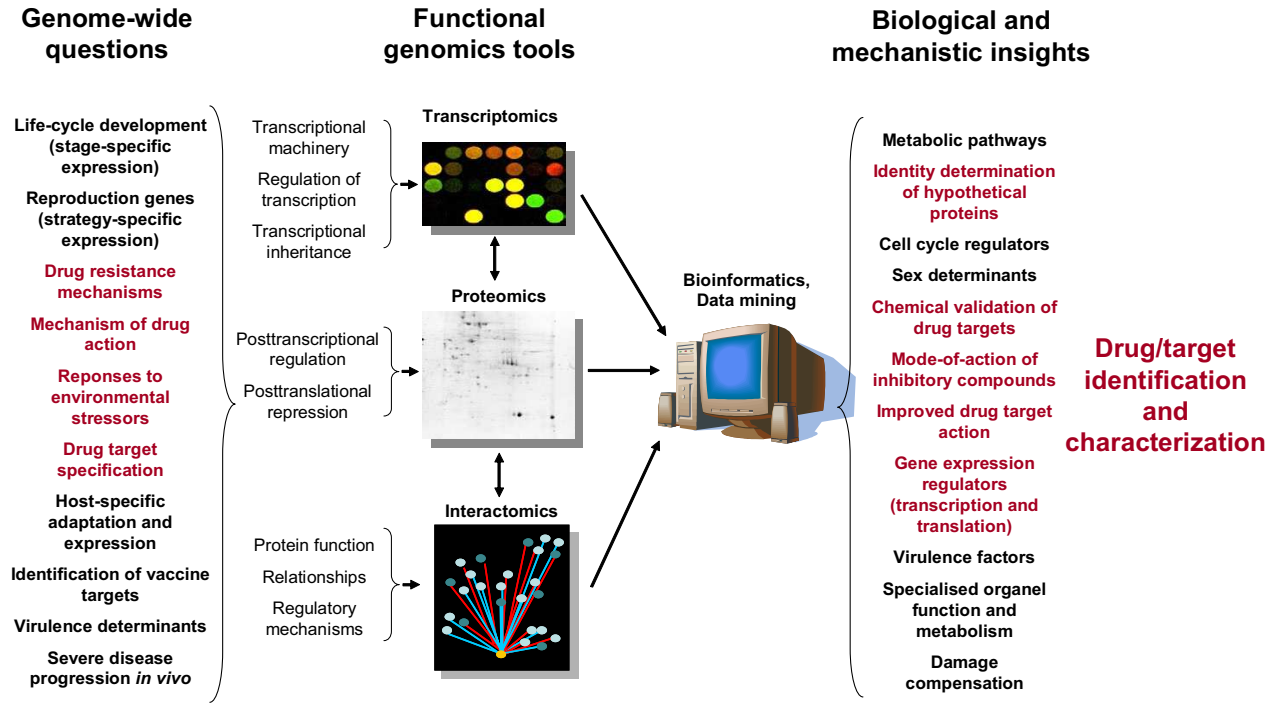

Figure 1

Malaria functional genomics (X-omics) strategies in the context of target and drug characterization. Selected questions that could be addressed by the application of functional genomics are listed, including those specific to the transcriptome, proteome or interactome (X-omes). Highlighted is the particular focus on the application of this type of strategies to drug(s) and target(s) characterization.

actively transcribed during the lifecycle but that 20% are specific of the intraerythrocytic developmental cycle and are produced in a periodic nature in a 'just-in-time' fashion. These early reports have been followed by investigations designed to answer numerous biological questions, including transcription and post-transcription specific aspects of the regulation of protein expression, transcriptional machinery and inheritance, interstrain conservancy, gametocytogenesis and antigenic variation control mechanisms [87-91]. A high degree of correlation exists between the *in vitro* and *in vivo* transcriptomes of *P. falciparum* with an overexpression seen for genes encoding a sexual stage antigen as well as gene families that encode surface proteins, providing interesting new vaccine candidates [92,93].

Proteomics studies are essential to conclusively prove mechanistic changes and explain global protein expression profiles, differential protein expression, posttranscriptional control, posttranslational regulation and modifications, alternative splicing and processing, subcellular localization and host-pathogen interactions (Figure 1). Reassuringly, there is a good correlation between the abundance of transcripts and the proteins encoded by these during the *P. falciparum* lifecycle [87,88] with a majority of discrepancies attributed to a delay between transcript production and protein accumulation. Analysis of the *P. falciparum* proteome [94] and a comprehensive and integrated analyses of the genome, transcriptome and proteome of *P. berghei* and *P. chabaudi chabaudi*, which represents the state-of-the-art of functional genomics applied to the lifecycles of *Plasmodia* [95,96], indicated that over half of the proteins in these parasites were detected solely in one stage of the lifecycle. This implies a considerable degree of specialization at the molecular level to support the demanding developmental pro-

gramme and suggested a highly coordinated expression of *Plasmodium* genes involved in common biological processes.

In-depth understanding of the protein-protein interaction network (defined as interactome) can provide insights into the function of proteins, regulatory mechanisms and functional relationships of these (Figure 1). An extensive protein-interaction study of *P. falciparum* (nearing global-type analysis) combining interaction information with co-expression data and GO annotations indicated unique interactions, identified groups of interacting proteins implicated in various biological processes, and predicted novel functions to previously uncharacterized proteins [97]. Interestingly, comparison of the *Plasmodium* interactome with that from other organisms indicated a marked divergence with very little conservation with other protein complexes [98].

The abovementioned X-ome datasets are therefore now available for *in silico* data mining approaches. As such, data mining of the transcriptome using an extensive sequence similarity search identified 92 putative proteases in the *P. falciparum* genome, 88 of which are actively transcribed [99]. This strategy has also been applied to the kinase family [100]. The transcriptome datasets are nowadays extensively utilized in PlasmoDB, MPMP and other specialized sites [101]. PlasmoDB additionally has links to proteome data and also allows access to the interactome data. Proteome experiments based on 2D-gel electrophoresis can furthermore be aided by Plasmo2D software to allow the identification of proteins in such platforms [102].

An important question in mining different global X-omic datasets is how they can be compared. Draghici et al. [103] made clear that the inconsistencies between the various microarray platforms (*in situ* synthesized short oligos, longer oligos, spotted oligos or cDNAs) are so high that it is almost impossible for the moment to compare results from different platforms. How can a transcriptomic profile be compared to a proteomic profile, given the errors, the linearity of signals and the magnitude of variations of each method, and the biological stability and turnover of RNA transcripts and of proteins? Is the enzyme profile correlated with the metabolite profile [104]? A pragmatic solution is to avoid the multiplicity of methods and by agreeing on some standards [103,105-107]. Although "global" invariant references might not exist, it is nevertheless worth trying to find in large gene expression matrices the most invariant (or less variant) genes [106]. This quest is of general concern and is currently one of the challenges proposed by the European Conferences on Machine Learning and the European Conferences on Principles and Practice of Knowledge Discovery in Databases [106,109]. In the absence of absolute references and standards, outputs from malaria X-omic experiments should be analysed and compared with caution, particularly when obtained with platforms that are distinct from those used to feed the public data repositories. Consequently, in addition to referential public repositories for functional genomics, versatile software for local analyses are strongly needed.

Various tools aim to provide biological interpretation of gene clusters but these mostly specialize in only one or two types of analyses. FatiGO [110], GeneLynx [111] and Gostat [112] are powerful tools for Gene Ontology mining; GoMiner [113], MAPPFinder [114] and DAVID [115] use GO and metabolic pathway interpretation whereas GeneXPress [116] and MiCoViTo [117] use metabolic pathways and incorporate transcription regulation visualization. Improvements on these include a web interface called MADIBA (MicroArray Data Interface for Biological Annotation, [118]) that has been initially designed for malaria transcriptomics (Dr. C. Claudel-Renard, personal communication). This interface links a relational database of various data sets to a series of analysis tools designed to facilitate investigations of possible reasons for co-expression of clusters of genes (e.g. from gene expression data) and to deduce possible underlying biological mechanisms. Clusters of co-expressed genes are automatically subjected to five different analytical modules including 1) search for over-represented GO terms in clusters, 2) visualization of related metabolic graphs with KEGG representations, 3) chromosomal localization, 4) search for motifs in the upstream sequences of the genes and finally 5) *Plasmodium*-specific genes without human homologues. MADIBA analysis of the transcriptome dataset from Le Roch et al. [85] resulted in an improved annotation of the *Plasmodium* genes (41% vs. 37%) and characterization of 6 additional clusters with GO annotations, of which one exclusively contained glycolysis in its entirety (except for fructose-bisphosphate aldolase) and another identifying gene as potential as drug targets due to their *Plasmodium*-specific characteristics. Therefore, MADIBA allows versatile analyses of a vast variety of transcriptomic profiles. These analyses can highlight potential drug targets by providing functionality to co-clustered expressed genes in a guilt-by-association manner, including those of un-annotated proteins, by predicting co-regulated expression *via* chromosomal localization as well as the identification of motifs for *cis*-regulatory elements and lastly by identifying unique *Plasmodium*-specific genes involved in specific biological mechanisms.

With the advent of integrative investigations of datasets from the transcriptome, proteome, interactome etc. new analysis tools are being developed including a Partial Least Squares (Projection to Latent Structures-PLS)

method which has been used to integrate yeast transcriptome and metabolome data [119]. Linear modelling was used to investigate the changes in the transcriptome due to environmental perturbations and, assuming that the metabolome is a function of the transcriptome, the metabolic variables were modeled with PLS. A genome-wide investigation of protein function was recently performed by computationally modelling the *P. falciparum* interactome [120] to elucidate local and global functional relationships between gene products. This novel approach entailed an integration of *in silico* and experimental functional genomics data within a Bayesian framework to create the network of pairwise functional linkages. This resulted in predicting functionality based on associations between characterized and uncharacterized proteins for 95% of the currently annotated hypothetical proteins in the *P. falciparum* proteome. Only 107 hypothetical proteins show interaction with other hypothetical proteins potentially representing new pathways or previously uncharacterized components of known pathways.

This overview shows that the integration and mining of global functional genomic experiments are in the frontline in drug discovery processes [121,122]. Malaria X-omics data provide comprehensive information to, amongst others, understand the mode-of-action of inhibitory compounds, allow optimization of drug action, validate drug targets (chemical validation strategies), identify families of genes/gene products that are more amenable as drug targets ('druggable genes'), annotate the function of hypothetical proteins by 'guilt-by-association' and point out specialized gene expression regulation systems (Figure 1) [123].

In the case of drug treatments, the sought effects on the metabolism of targeted tissues or organisms include up- or downregulation of the protein target(s), the upregulation of detoxification pathways (cytotoxic responses) and the upregulation of alternative or compensatory pathways of the affected organism that can be reflected in changes in the transcriptome/proteome of the organism. The characterization of the transcriptional response induced by drug challenge has been applied with success in the antibacterial field, creating reference compendia of expression profiles after drug challenge that provide insight into a drugs' MOA [123-125]. Transcriptional profiling of drug challenged malaria parasites has been limited to only a few studies to date including Serial analysis of gene expression (SAGE) of chloroquine treated parasites [126], a high-density short-oligonucleotide array study on parasites treated with phosphatidylcholine biosynthesis inhibitors [88] and custom arrays originating from suppression, subtractive hybridisation (SSH) libraries on parasites treated with polyamine biosynthesis inhibitors [127]. Drug-specific transcriptional responses were seen in the chloroquine and polyamine inhibition studies indicating the presence of a feedback signaling mechanism. Le Roch et al. [88] compared transcriptional responses with proteome analyses and showed that more pronounced changes were induced at the protein level after drug challenge. This is also true for antifolate inhibited parasites where a marked increase were seen in folate biosynthesis protein levels upon treatment with inhibitors against DHRF-TS [128]. Global-level proteome response analysis of the combination of artemether and lumefantrine also revealed drug-specific changes in the proteome [129]. Subproteomic investigations have additionally become particularly important to determine the molecular binding partner/target protein of an inhibitory compound and/or to describe the mode-of-action of such compounds. This has been applied to ferriprotoporphyrin IX were identified [130], kinase inhibitors [131] and the quinoline family of compounds [132].

Any of the above-mentioned strategies are potential starting points to the discovery of unsuspected drug targets in *P. falciparum*, whether it is a new/additional functionality that is ascribed to a known protein or the characterization of novel function of a previously 'unknown' protein. As the hypothetical proteins represent more than half (~60%) of the malarial proteome (see above), these are some of the most attractive areas to the drug target discovery effort. The basic rationale behind using expression profiles to assign functionality to genes is based on the principle of guilt-by-association [133,134] in which genes coding for proteins with similar functionality often exhibit the same expression profiles and protein-interacting partners. Coincidentally, the expression profiles of genes specific to a given organelle also display similar expression patterns. This principle has been applied with unsupervised robust k-means clustering [85]. Improvements on this clustering approach were proposed using a semi-supervised clustering method called ontology-based pattern identification (OPI) [135]. OPI uses previous gene annotation data to generate clusters with greater specificity and confidence whose members then additionally share the same expression profiles. However, Llinás and del Portillo [136] warns against using only classical guilt-by-association methods, showing that many genes that are functionally unrelated show similar expression profiles during the asexual development of *P. falciparum*.

Malaria protein structures and virtual ligand screening on candidate targets

Vital malaria proteins may have no counterpart in humans or sequence dissimilarity with their human homologues that may be sufficient to become a therapeutical target without disturbing essential function in the human host. The literature on potential protein targets for anti-malaria treatments is already crowded, and will

hopefully be enriched and better documented in the future. Once a target has been identified, an important decision is whether or not to enter into costly drug- or vaccine-development programmes. Determination or prediction of the target three-dimensional structure and *in silico* experiments can be achieved to connect these biological targets with the pharmacological space of small compounds [137], and assist this decision and provide initial clues on possible therapeutical strategies.

Malaria protein structures

The rationalized identification of new inhibitors depends on possession of structural information. As for any other organism, the primary problem is obtaining high and pure protein yields for crystallization trials. Recombinant expression of malarial proteins in *E. coli* is notoriously difficult, however. A number of problems are typically encountered. The A+T richness results in substantially different codon usage compared to *E. coli*. *Plasmodium* genes are also typically much longer than their homologues in other organisms, as are the resulting proteins. Increased protein size is due mostly to long protein inserts with generally little homology to cognate enzymes. These inserts tend to be disordered and of low complexity, resulting in proteins that are not amenable to expression and crystallization. Further problems include sporadic mutations of low complexity sequences introduced by *E. coli*, and cryptic prokaryotic translation start sites within malarial genes. Some level of protein expression may be obtained by fine control of expression conditions, often a change of strain or of complete expression system, addition of rare codon tRNAs, and more and more often by production of synthetic genes coding for identical protein sequences but with a codon usage optimized for bacteria [138-142]. Mehlin et al. [143] recently attempted a wholesale expression of 1000 malarial genes and obtained soluble expression for only 63 genes. High predicted disorder, molecular weight, pI and lack of homology to *E. coli* proteins were all negatively correlated with soluble expression.

The difficulty of expressing malarial proteins is reflected by the paucity of structures in the Protein Data Bank [144]. At the time of writing there are only 64 non-redundant *Plasmodium* protein structures in the PDB [see Additional file 1]. In contrast, querying the PDB for human entries (excluding > 90% sequence identity) reveals more than 1700 structures. The advent of structural genomics programmes (the Structural Genomics Consortium, [145]; the Structural Genomics of Pathogenic Protozoa, [146]) has increased the throughput of new malarial structures. Since 2003, 56 depositions of *Plasmodium* structures have been made. Whether this trend will continue beyond the "low hanging fruit" remains to be seen.

In lieu of crystal structures for malarial proteins many groups have resorted to homology modelling. This approach depends critically on the alignment with template structures. Unfortunately the biased nucleotide and amino acid composition (see above and [58]) and *Plasmodium*-specific inserts make it difficult to correctly identify core-conserved regions. The presence of inserts often confuses multiple and structural-alignment programmes. A number of techniques have been used to circumvent this problem (Figure 2). From a first pass alignment, approximate insert positions can be determined. Sequences can then be split according to long inserts and re-aligned. Inserts can vary considerably across different *Plasmodium* species ([147] and C. Claudel-Renard, personal communication). While adjusting an alignment for modelling, it is useful to refer to phylogenetically diverse multiple alignments including as many *Plasmodium* sequences as possible (see above, [148]). As an adjunct to alignment, independent motif identification (*e.g.* the MEME system; [149,150]) can be used to fix mistakes that alignment programmes frequently make when aligning long *Plasmodium* proteins with homologues [148,151]. Further improvements can be made by using hydrophobic cluster analysis [61] and secondary structure predictions to align homologous regions within inserts. Once an alignment has been decided on, based on visual assessment, a series of models can be built. Because of the high degree of uncertainty that often accompanies alignments used for modelling malarial proteins, it is usually not feasible to rectify all structural anomalies. But by performing standard quality checks on a large sample of models and summarizing the results, it is possible to identify parts of the alignment causing most problems. Refined alignments might benefit from species-specific matrices that take into account the differences of amino acid distribution between the aligned proteins [60,152].

Despite the difficulties with homology modelling of malarial proteins there have been some notable successes. Malarial DHFR forms part of a bifunctional protein that also carries thymidylate synthase. A number of existing drugs such as cycloguanil and pyrimethamine target the DHFR domain, and have been used effectively in the past. However drug resistance has evolved that reduces the usefulness of this important class of drugs. Hence malarial DHFR has been a popular target for homology modelling efforts [153-158]. Toyoda et al. [153] were able to identify new inhibitors in the micromolar range. McKie et al. [154] and Lemcke et al. [155] could rationalise the pyrimethamine resistance caused by the S108N mutation. One of these models was further used to identify new inhibitors acting in the nano- and micromolar ranges [154]. Delfino et al. [158] in turn used their model to investigate a large number of antifolate resistant mutants. Rastelli et al. [156] further explained the cycloguanil resistance/

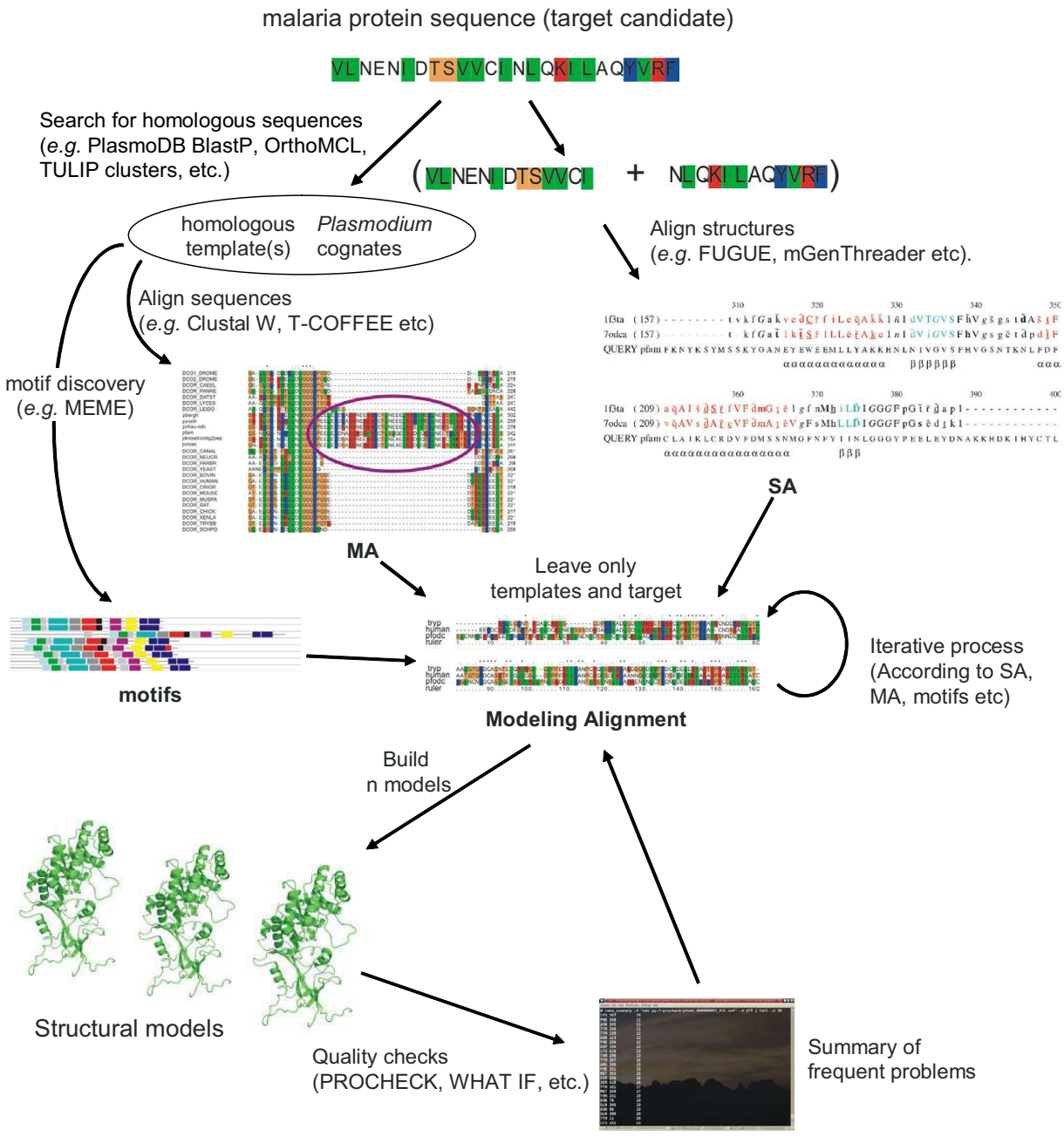

Figure 2

Current pipeline for the homology-based modelling of malaria protein 3D-structures. This scheme emphasizes on the currently available methods to overcome amino acid bias, low sequence identity, protein inserts etc. Future upgrades include the refinement of each of these methods, for instance implementing asymmetric substitution matrices discussed in the text, that take into account the different amino acid distributions of malarial and non-malarial proteins for pairwise alignments.

pyrimethamine sensitivity conferred by A16V+S108T, as well as the ability for WR99210 to inhibit both pyrimethamine and cycloguanil resistant mutants. A number of new inhibitors were also successfully designed. The high accuracy of the alignment used for modelling meant that predicted dockings were subsequently confirmed with the crystal structure of the complete bifunctional enzyme [159]. Considerable work has also gone into modelling malarial proteases essential to the parasite's intra-erythrocytic life stage. A number of these models have been used to identify new inhibitors [160-163], although the increasing number of crystal structures for these proteases is likely to gradually replace the need for homology models [see Additional file 1].

Apart from the PDB, there is currently no resource that includes all *Plasmodium* protein structures. Both publicly available versions of PlasmoDB (4.4 and 5) are still incomplete in this regard. PlasmoDB 4.4 only lists *P. falciparum* structures, and there is incomplete overlap between versions 4.4 and 5. PlasmoDB 4.4 includes a section for >440 modeled structures based on a wholesale modelling attempt of >5,000 open reading frames documented at [164]. However, at the time of writing the actual structures were not available. Furthermore, PlasmoDB 5 also lists non-*Plasmodium* structures therefore requires some curating. A dedicated resource for deposition of structures from the *Plasmodium* structural community would be useful. The resource should also include model structures subjected to the same rigor as experimental structures in the PDB, including quality criteria and scoring. Thus "experimental" details for modelling (alignments, software methods etc) should be included. This resource should also allow easy comparison with corresponding solved structures enabling evaluation of the communities' ability to model this difficult organism.

Protein structures provide invaluable information for drug or vaccine discovery. First, compliance of the structure with known properties of recorded pharmaceutical targets (termed "druggable" genes) [165-167] or prediction of the occurrence of immunogenic epitopes (termed "immunizable" genes) [168,169], can be investigated. Second, having a protein structure in hand is a very early, but necessary step, for the *in silico* prediction of the families of ligands that can interfere with protein function.

Virtual ligand screening

Advances in combinatorial chemistry have broken limits in organic synthetic chemistry and accelerated the production throughput. Thus, millions of chemical compounds are currently available in private and academic laboratories and recorded in 2D or 3D electronic databases. It is often technically impossible and very expensive to screen such a high number of compounds using wet high throughput screening techniques. An alternative is high throughput virtual screening by molecular docking, a technique which can screen millions of compounds rapidly and cost effectively. Molecular docking is a computer-based method which predicts the ligand conformations inside the active site of the target as well as an estimate of the binding affinity between protein and ligand. It also gives insight about the interactions between protein and ligand and allows to generate mode-of-action hypotheses. Screening each compound, depending on its structural complexity, requires from a few seconds to hours of computation time on a standard PC workstation depending on the chosen docking algorithm. Consequently, screening all compounds in a single database would require years. However, the problem is embarrassingly parallel and the computation time can be reduced very significantly by distributing data to process over a grid gathering thousands of computers [170,171].

Recently, virtual screening projects on grids have emerged with the purpose of reducing cost and time. They focused on the development of an *in silico* docking pipeline on grids of clusters [172] but also on the optimization of molecular modelling [173]. Pharmaceutical laboratories have also become interested by the grid concept; Novartis deployed the first automated modelling and docking pipeline on an internal grid [174]. Other projects focused on virtual screening deployment on a pervasive grid, or desktop grid, to analyse specific targets [175].

In mid-2005, the WISDOM (World-wide In Silico Docking On Malaria) initiative successfully deployed large scale *in silico* docking on the European public EGEE grid infrastructure [176]. The biological targets were plasmepsins, aspartic proteases of *Plasmodium* responsible for the initial cleavage of human haemoglobin [177]. There are ten different plasmepsins coded by ten different genes in *P. falciparum* (Plm I, II, IV, V, VI, VII, VIII, IX, X and HAP) [178]. High levels of sequence homology are observed between different plasmepsins (65–70%). Simultaneously they share only 35% sequence homology with their nearest human aspartic protease, Cathepsin D4 [179]. This and the presence of accurate X crystallographic data made plasmepsin an ideal target for rational drug design against malaria.

The main goal of the WISDOM project has been to make use of the EGEE grid infrastructure to set up an *in silico* experimentation environment. Having the necessary computing power at hands, scientists from a virtual organization can design new large-scale test systems for generating new hypotheses. The benefit is that a large number of targets can be combined with a very large number of potential hit molecules, using different docking algorithms and allowing to chose different parameter settings. As dis-

cussed above for the X-omic experiments in post-genomic platforms, *in silico* methods pose the same important question in order to mine the data, *i.e*. how they can be compared. The careful input data preparation is a crucial step in the process, which has to be performed by experts and be made available for the whole scientific community. By sharing the results of virtual screenings in a common knowledge space, different experts coming from different fields can jointly derive a rational for selecting appropriate combinations of targets, ligands and virtual screening methods.

The WISDOM large scale *in silico* docking deployment saw over 46 million docked ligand-protein solutions, resulting from two docking tools, five targets, one million compounds and four parameter settings, the equivalent of 80 years on a single PC in about six weeks. Up to 1,700 computers were simultaneously used in 15 countries around the world. Post-processing of the huge amount of data generated was a very demanding task as millions of docking scores had to be compared. At the end of the large scale docking deployment, the best compounds were selected based on the docking score, the binding mode of the compound inside the binding pocket and the interactions of the compounds to key residues of the protein.

Several promising scaffolds have been identified among the 100 compounds selected for post processing. Among the most significant ones are urea-, thiourea-, and guanidino analogs, as these scaffolds are most repeatedly identified in the top 1,000 compounds (Figure 3). Validating this approach, some of the compounds identified were similar to already known plasmepsin inhibitors, like urea analogs from the Walter Reed chemical database, which were previously established as micro molar inhibitors for plasmepsins [180]. This indicates that the overall approach is sensible and large scale docking on computational grids has real potential to identify new inhibitors. In addition, guanidino analogs appeared very promising and most likely to become a novel class of plasmepsin inhibitors.

The developed and established protocols can be used for coming scenarios. Several teams have expressed interest to propose targets for a second computing challenge called WISDOM II that should be carried out in late 2006. While docking methods have been significantly improved in the last years, docking results need to be post-processed with more accurate modelling tools before biological tests are undertaken. The major challenges for docking methods are prediction and scoring. Molecular dynamics (MD) has great potential at this stage: firstly, it enables a flexible treatment of the ligand/target complexes at room temperature for a given simulation time, and therefore is able to refine ligand orientations by finding more stable complexes; secondly, it partially solves conformation and orientation search deficiencies which might arise from docking; thirdly, it allows the re-ranking of molecules based on more accurate scoring functions. Efforts are now devoted to deploy on the grid both docking and Molecular Dynamics calculations to further accelerate *in silico* virtual screening before in vitro testing.

Conclusions: toward a chemogenomic knowledge space

In this paper, five aspects of the *in silico* storage, organization and mining of data from malaria genomics and post-genomics, were examined in the context of the prediction and characterization of targets and drugs: 1) the comparison of protein sequences including compositionally atypical malaria sequences, 2) the high throughput reconstruction of molecular phylogeny, 3) the representation of biological processes particularly metabolic pathways, 4) the versatile methods to integrate genomic data, biological representations and functional profiling obtained from X-omic experiments after drug treatments, 5) the determination and prediction of protein structures and their virtual docking with drug candidate structures. Data management that should be at the genomic scale, including multiple species, and should therefore be as reliable as possible. A "biological space" should be formatted so as to represent scientific knowledge and to connect genomic data and post-genomic functional profiles in the most versatile way, allowing the mining of information with diverse methods (Figure 4a–e). This "biological space" should be linked to a "chemical space" that contains all small molecules stored in chemolibraries (millions of compounds) including known drugs (Figure 4f–h). Thus, progresses toward a chemogenomic knowledge space will benefit on the referential public repositories and particularly UniProt and PlasmoDB, on recent theoretical advances in genomics and post-genomics data management and mining and on the power of computer grids (Figure 4).

Genomic data (Figure 4a), *i.e*. protein sequences, can be organized based on sequence similarity (Figure 4b). This projection of protein sequences should allow the high throughput reconstruction of molecular phylogenies both at the intraspecific (connecting paralogs and alleles) and interspecific (connecting homologues among which orthologs) levels, following statistically accurate methods (Figure 4b). This task is difficult because of the compositional bias and high insert content of malaria sequences. Statistically valid protein sequence comparisons are now available to allow genome scale alignment and high throughput phylogeny reconstruction (named TULIP), including malaria atypical sequences, and providing quality scores on which one can rely after automatic genome scale treatments. Benefits from the TULIP method include an easy upgrade and update of the obtained protein spa-

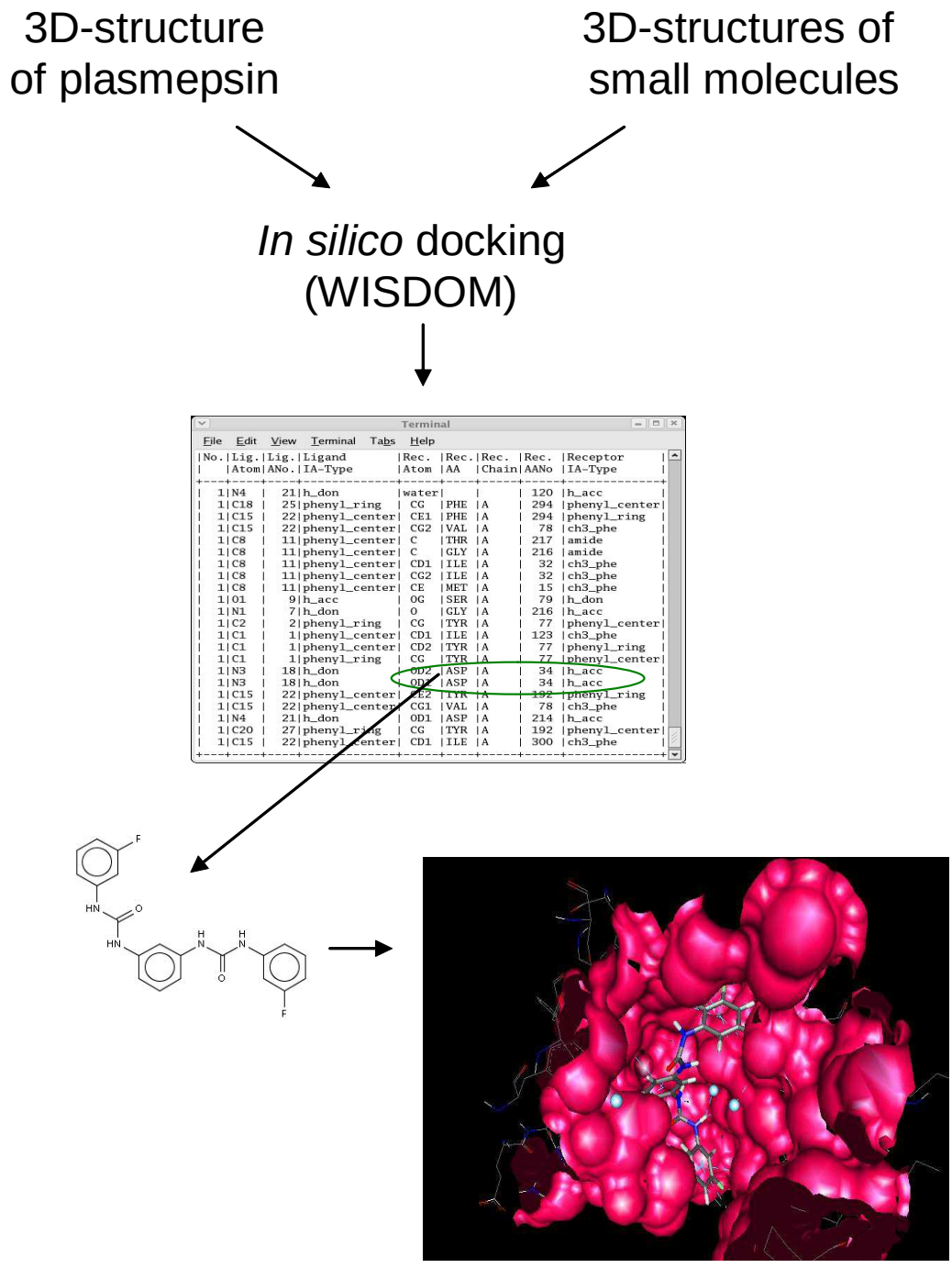

Figure 3

***In silico* screening for protein ligands based on structural docking**. A urea analog inhibiting malaria plasmepsins was identified with good score from the first WISDOM (World-wide In Silico Docking On Malaria) campaign. The WISDOM initiative successfully deployed large scale *in silico* docking on the European public EGEE grid infrastructure. The ligand shown here docks inside the binding pocket of plasmepsin, and interacts with key protein residues. The developed and established protocols can be used for new targets, and particularly a second computing challenge, WISDOM II.

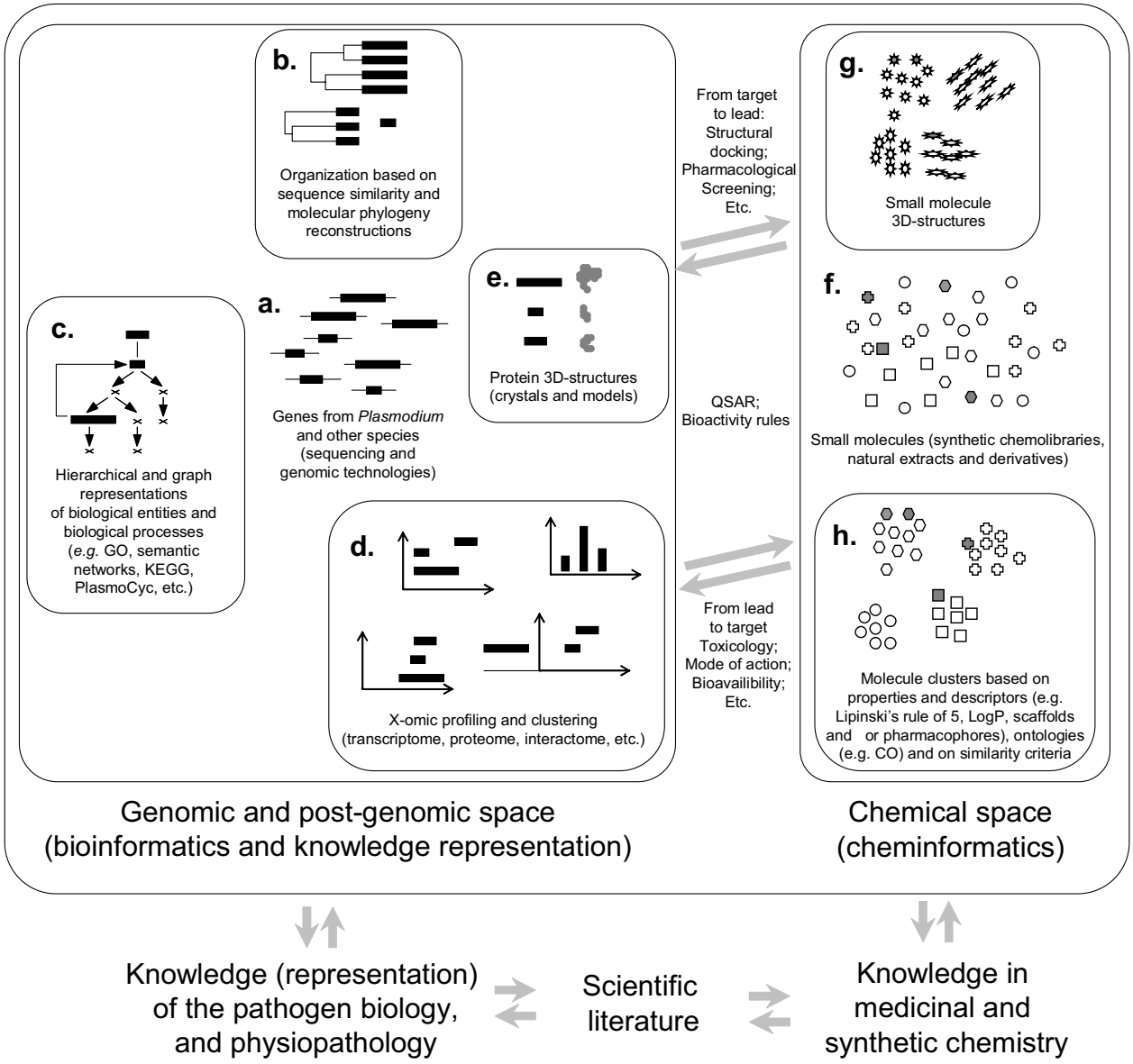

Figure 4
Malaria chemogenomics: organization and treatment of genomic, post-genomic and chemical information for the prediction and characterization of target and drug candidates. Genomic data from *Plasmodium* and other species (a), *i.e.* protein sequences, should be organized based on sequence similarity (b). This projection should allow the high throughput reconstruction of molecular phylogenies both at the intraspecific (connecting paralogs and alleles) and interspecific (connecting homologues among which orthologs) levels following statistically accurate methods *e.g.* the TULIP method. Another substantial side of the biological space is designed by representing the knowledge of the biological processes, using stable ontologies *e.g.* the GO, and dynamic graph representation, *e.g.* PlasmoCyc (c). Versatile tools should allow the integration of genomic data, biological process representations and global functional profiles obtained with diverse X-omic approaches (4). These tools should comply with the large diversity of technologies and mining methods. The collection of information on the biological response to drugs is one of the doors to connect the biological space with the chemical space, following the "reverse chemical genetic" way, *i.e.* "from known drugs to biological response" (toxicity, mode of action). The other door to connect the chemical space and the biological space follows the "direct chemical genetic" way, *i.e.* "from known biological target to drug candidates". In addition to malaria protein structures obtained from crystals, the automated structural annotation of the malaria proteome should be initiated with quality scores (e). Based on protein structure information, virtual docking campaigns such as the WISDOM challenges can be achieved using the power of computer grids. The *in silico* organization of the small molecules stored in chemolibraries (f) follows similar principles, in particular the determination of three-dimensional structures of small molecules (g) and a clustering of small molecular structures based on drug properties and descriptors (h). Sharing and mining of chemogenomic information, completed with knowledge harvested in unstructured scientific literature, would benefit of the advances in knowledge space design and deployment on knowledge grids.

tial projection. Another substantial side of a biological space contains representations of the knowledge of biological processes, using stable ontologies, and dynamic graph representation (Figure 4c). Here, efforts are still needed to improve existing representations of the metabolism and numerous projects are under progress, most importantly PlasmoCyc and PMPM. The PlasmoCyc existing metabolic graphs have the advantage of being more easily updated and usable for *in silico* mining methods, as long as the output is examined by biological experts. In a brief overview of malaria X-omic profiling, some tools allowing a linkage between genomic data, biological process representations and global functional profiles have been introduced (Figure 4d). There is still a large diversity of data treatment and mining strategies, reflecting the diversity of technologies and mining methods. This step is one of the doors to connect the biological space with the chemical space, following the "reverse chemical genetic" way, *i.e.* "from known drugs to biological response" (toxicity, mode of action). Basic analytical tools like MADIBA and sophisticated mining approaches will be needed to understand and compare the biological responses to antimalarial drugs. The other door to connect the biological space and the chemical space follows the "direct chemical genetic" way, *i.e.* "from known biological target to drug candidate". In addition to crystal structures of malaria proteins, the automated structural annotation of the malaria proteome should be initiated (Figure 4e). Based on protein structure information, virtual docking campaigns such as the WISDOM challenges can be achieved using the power of computer grids.

This paper did not review the "chemical" side for chemogenomics space. By numerous aspects the *in silico* organization of the small molecules stored in chemolibraries (Figure 4f) was not achieved the way biological information was. Clustering of small molecular structures based on properties is highly debated (Figure 4h). Collections of small molecules are generally designed to obey the pragmatic Lipinski's "rule of five" [181] making them likely candidates for drug discovery. Numerous studies are under progress in order to identify which small molecule descriptors can comply with chemogenomic approaches (*e.g.* the Accamba project for the analysis of chemolibraries and the building of bioactivity models; [182]). A chemical ontology (CO) has been recently introduced [183], but it has not yet been used and validated by the scientific community the way the GO was. A database for Chemical Entities of Biological Interest (ChEBI) has also been launched [184]. The modelling of the three-dimensional structures of small molecules (Figure 4g) can be predicted by numerous public or commercial methods which should be examined with attention (see the WISDOM challenges). PubChem, a repository for molecules acting on biological targets was recently launched [183,185] and the UniProt protein knowledge base was recently upgraded to report toxic doses of small molecules on proteins [186,187], however these initiatives are just starting points. Access to an ocean of small molecular structures and to a deluge of biological sequences raises an enthusiastic challenge: "The goal for the coming decades will be to explore the overlap between chemistry space and protein space" [188]. Is this prediction exuberant? From the methodological survey summarized in this paper, this next milestone for malaria research is not out of reach.

Beyond virtual screening, the grid technology provides the collaborative IT environment to enable the coupling between molecular biology research and goal-oriented field work [189]. It proposes a new paradigm for the collection and analysis of distributed information where data do no longer need to be centralized in one single repository. On a grid, data can be stored anywhere and still be transparently accessed by any authorized user. The computing resources of a grid are also shared and can be mobilized on demand so as to enable very large-scale genomics comparative analysis and virtual screening. A longer term perspective is, therefore, to enhance the ability to share diverse, complex and distributed information on a given disease for collaborative exploration and mutual benefit. The concept of a knowledge space is to organize the information so that it can be reached in a few clicks. This concept is already successfully used internally by pharmaceutical laboratories to store knowledge [175]. The grid permits the building of a distributed knowledge space so that each participant is able to keep the information he owns on his/her local computer. A set of grid services would particularly take advantage of the developments in the area of semantic text analysis for extraction of information in biology and genome research. (Figure 4, lower part). The literature on malaria biology, physiopathology and medicinal chemistry is, at least in significant parts, stored as unstructured texts that make an invaluable source of knowledge which access depends on advances in terminology analysis and term extraction methods. A first attempt at using terminology analysis for the "harvesting" of relevant concepts in a defined disease area has been undertaken in another biomedical grid project, the recently started EU Integrated Project @neuRIST [190]. In this project, terminology analysis lead to the refinement of a disease-specific text corpus and provided a shortlist of relevant terms. Moreover, not only the genes associated with the defined disease area could be identified by automated methods, but also single nucleotide polymorphisms (SNPs) published for these disease-associated genes, could be automatically identified (Dr. Laura Furlong, IMIM, Barcelona, personal communication). An important focus of future activities in this project is the construction and validation of fine-

granular disease-specific ontologies. This concept can easily be adopted for a knowledge base on any disease, including malaria. Existing databases can be complemented by this automatically generated semantic layer, which subsequently would also be helpful for data mediation. Moreover, a structured knowledge space would produce grid services for indexing of distributed data resources and thus improve navigation through knowledge and retrieval of relevant information.

Finally, present and future *in silico* information must be supported and validated by data gathered *in vitro* and *in vivo*. As other X-omic strategies approach, the necessity of *in silico* mining is unquestionable; it is also susceptible to generate a huge amount of theoretical data that will need years to be confronted to the experimentation. The challenges in this domain remain severe. Difficulties of cloning and expressing parasite proteins in heterologous systems, the validation of druggable targets using siRNA etc, and the experimental assignment of functionality to the many hypothetical proteins, will occupy scientists in parallel for some time to come. Data generated from *in silico* analysis leads to a need for further laboratory work, such as the *in vitro* testing of ligands identified as potential drug leads through the WISDOM project. A strong linkage between scientists undertaking *in vitro* research and *in silico* researchers is therefore essential to support an iterative approach to knowledge generation and analysis, in the context of malaria chemogenomics.

Authors' contributions

All authors contributed to the analysis of the current status on the *in silico* storage and organization of malaria genomics, post-genomics data and antimalaria chemical information, as made available in the literature and public websites. All authors contributed to the writing of the current review manuscript.

Additional material

Additional File 1

Non-redundant malarial structures in the Protein Data Bank (PDB). The table compiles non redundant entries for Plasmodium *protein structures in the Protein Data Bank at the date of writing.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1475-2875-5-110-S1.doc]

Acknowledgements

We are indebt to Dr. Jane Morris for discussions and in-depth reading of this manuscript. We thank Dr. Henri Vial for fruitful discussions, and Dr. Clotilde Claudel-Renard and Dr. Laura Furlong for sharing personal information. This publication was made possible through financial supports from the South African National Research Foundation, the South African National Bioinformatics Network, the French Agence Nationale de la Recherche (PlasmoExplore project), Région Rhône-Alpes (Cluster 9, E.M. and Prospective, N.S.), Fondation pour la Recherche Médicale (O.B.) and French Ministère des Affaires Etrangères. EGEE-II is a project funded by the European Union under contract number INFSO-RI-031688. This publication was further supported by the New Partnership for Africa's Development (NEPAD), African Union.

References

1. **World Malaria Report 2005.** *Geneva, World Health Organization, WHO/UNICEF* 2005.
2. Grover-Kopec EK, Blumenthal MB, Ceccato P, Dinku T, Omumbo JA, Connor SJ: **Web-based climate information resources for malaria control in Africa.** *Malar J* 2006, **5:**38.
3. Adl SM, Simpson AG, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S, James TY, Karpov S, Kugrens P, Krug J, Lane CE, Lewis LA, Lodge J, Lynn DH, Mann DG, McCourt RM, Mendoza L, Moestrup O, Mozley-Standridge SE, Nerad TA, Shearer CA, Smirnov AV, Spiegel FW, Taylor MF: **The new higher level classification of eukaryotes with emphasis on the taxonomy of protists.** *J Eukaryot Microbiol* 2005, **52:**399-351.
4. Desowitz RS: **Malaria: from quinine to the vaccine.** *Hosp Pract* 1992, **27:**209-14. 217–24, 229–32
5. Utzinger J, Tanner M, Kammen DM, Killeen GF, Singer BH: **Integrated program is key to malarial control.** *Nature* 2002, **419:**431.
6. Baldwin PC: **How night air became good air, 1776–1930.** *Environmental History* 2003, **8.3:**36 pars.
7. Nchinda T: **Malaria: a reemerging disease in Africa.** *Emerg Infect Dis* 1998, **4:**398-403.
8. Ridley RG: **Malaria: dissecting chloroquine resistance.** *Curr Biol* 1998, **8:**R346-R349.
9. Ridley RG: **Medical need, scientific opportunity and the drive of antimalarial drugs.** *Nature* 2002, **415:**686-693.
10. Farooq U, Mahajan RC: **Drug resistance in malaria.** *J Vector Borne Dis* 2004, **41:**45-53.
11. Baird JK: **Effectiveness of antimalarial drugs.** *N Engl J Med* 2005, **352:**1562-1577.
12. Waters A: **Malaria: new vaccines for old?** *Cell* 2006, **124:**689-693.
13. Jambou R, Legrand E, Niang M, Khim N, Lim P, Volney B, Ekala MT, Bouchier C, Esterre P, Fandeur T, Mercereau-Puijalon O: **Resistance of Plasmodium falciparum field isolates to in-vitro artemether and point mutations of the SERCA-type PfATPase6.** *Lancet* 2005, **366:**1960-1963.
14. Towie N: **Malaria breakthrough raises spectre of drug resistance.** *Nature* 2006, **440:**852-853.
15. Afonso A, Hunt P, Cheesman S, Alves AC, Cunha CV, do Rosario V, Cravo P: **Malaria parasites can develop stable resistance to artemisinin but lack mutations in candidate genes atp6 (encoding the sarcoplasmic and endoplasmic reticulum Ca2+ ATPase), tctp, mdr1, and cg10.** *Antimicrob Agents Chemother* 2006, **50:**480-489.
16. Hoffman SL, Bancroft WH, Gottlieb M, James SL, Burroughs EC, Stephenson JR, Morgan MJ: **Funding for malaria genome sequencing.** *Nature* 1997, **387:**647.
17. Gardner MJ: **The genome of the malaria parasite.** *Curr Opin Genet Dev* 1999, **9:**704-708.
18. Carucci DJ, Goodwin PM, Gottlieb M, McGovern V: **The *Plasmodium falciparum* genome project.** In *Malaria parasites: genome and molecular biology* Edited by: Waters AP, Janse CJ. Caister Academic Press, England; 2004:1-6.
19. Hall N, Gardner M: **The genome of *Plasmodium falciparum*.** In *Malaria parasites: genome and molecular biology* Edited by: Waters AP, Janse CJ. Caister Academic Press, England; 2004:7-31.
20. Carlton J, Silva J, Hall N: **The genome of model malaria parasites, and comparative genomics.** In *Malaria parasites: genome and molecular biology* Edited by: Waters AP, Janse CJ. Caister Academic Press, England; 2004:33-63.
21. Kooij TWA, Janse CJ, Waters AP: ***Plasmodium* post genomics: better the bug you know?** *Nature Rev* 2006, **4:**344-356.
22. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shal-

lom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B: **Genome sequence of the human malaria parasite *Plasmodium falciparum*.** *Nature* 2002, **419:**498-511.

23. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, Peterson JD, Pop M, Kosack DS, Shumway MF, Bidwell SL, Shallom SJ, van Aken SE, Riedmuller SB, Feldblyum TV, Cho JK, Quackenbush J, Sedegah M, Shoaibi A, Cummings LM, Florens L, Yates JR, Raine JD, Sinden RE, Harris MA, Cunningham DA, Preiser PR, Bergman LW, Vaidya AB, van Lin LH, Janse CJ, Waters AP, Smith HO, White OR, Salzberg SL, Venter JC, Fraser CM, Hoffman SL, Gardner MJ, Carucci DJ: **Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*.** *Nature* 2002, **419:**512-519.
24. Duraisingh M, Ferdig MT, Stoeckert CJ, Volkman SK, McGovern VP: ***Plasmodium* research in the postgenomic era.** *Trends Parasitol* 2006, **22:**1-4.
25. Roos DS: **Themes and variations in apicomplexan parasite biology.** *Science* 2005, **309:**72-73.
26. Chaudhary K, Roos DS: **Protozoan genomics for drug discovery.** *Nat Biotechnol* 2005, **23:**1089-1091.
27. Coppel RL: **Bioinformatics and the malaria genome: facilitating access and exploitation of sequence information.** *Mol Biochem Parasitol* 2001, **118:**139-145.
28. Kissinger JC, Brunk BP, Crabtree J, Fraunholz MJ, Gajria B, Milgram AJ, Pearson DS, Schug J, Bahl A, Diskin SJ, Ginsburg H, Grant GR, Gupta D, Labo P, Li L, Mailman MD, McWeeney SK, Whetzel P, Stoeckert CJ, Roos DS: **The *Plasmodium* genome database.** *Nature* 2002, **419:**490-492.
29. Bahl A, Brunk B, Crabtree J, Fraunholz MJ, Gajria B, Grant GR, Ginsburg H, Gupta D, Kissinger JC, Labo P, Li L, Mailman MD, Milgram AJ, Pearson DS, Roos DS, Schug J, Stoeckert CJ, Whetzel P: **PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data.** *Nucleic Acids Res* 2003, **31:**212-215.
30. Stoeckert CJ Jr, Fischer S, Kissinger JC, Heiges M, Aurrecoechea C, Gajria B, Roos DS: **PlasmoDB v5: new looks, new genomes.** *Trends Parasitol* in press.
31. Gene Ontology Consortium: **The Gene Ontology (GO) project in 2006.** *Nucleic Acids Res* 2006, **34:**D322-D326.
32. Heiges M, Wang H, Robinson E, Aurrecoechea C, Gao X, Kaluskar N, Rhodes P, Wang S, He CZ, Su Y, Miller J, Kraemer E, Kissinger JC: **CryptoDB: a *Cryptosporidium* bioinformatics resource update.** *Nucleic Acids Res* 2006, **34:**D419-D422.
33. Martin DM, Berriman M, Barton GJ: **GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes.** *BMC Bioinformatics* 2004, **5:**178.
34. Bastien O: **Theoretical advances and numerical methods for genomes comparisons. Application to the *Plasmodium falciparum*/*Arabidpsis thaliana* genomes and proteomes comparison.** In *PhD thesis* Grenoble University; 2006.
35. Cavalier-Smith T: **Kingdom protozoa and its 18 phyla.** *Microbiol Rev* 1993, **57:**953-994.
36. Archibald JM, Keeling PJ: **Recycled plastids: a 'green movement' in eukaryotic evolution.** *Trends Genet* 2002, **18:**577-584.
37. McFadden GI, Reith ME, Munholland J, Lang-Unnasch N: **Plastids in human parasites.** *Nature* 1996, **381:**482.
38. Köhler S, Delwiche CF, Denny PW, Tilney LG, Webster P, Wilson RJ, Palmer JD, Roos D: **A plastid of probable green algal origin in Apicomplexan parasites.** *Science* 1997, **275:**1485-1489.
39. Soldati D: **The apicoplast as a potential therapeutic target in and other apicomplexan parasites.** *Parasitol Today* 1999, **15:**5-7.
40. Roos DS: **The apicoplast as a potential therapeutic target in *Toxoplasma* and other apicomplexan parasites: some additional thoughts.** *Parasitol Today* 1999, **15:**41.
41. Maréchal E, Cesbron-Delauw MF: **The apicoplast: a new member of the plastid family.** *Trends Plant Sci* 2001, **6:**200-205.
42. Waller RF, McFadden GI: **The apicoplast: a review of the derived plastid of apicomplexan parasites.** *Curr Issues Mol Biol* 2005, **7:**57-79.
43. Bisanz C, Botté C, Saïdani N, Bastien O, Cesbron-Delauw MF, Maréchal E: **Structure, function and biogenesis of the secondary plastid of apicomplexan parasites.** In *Current Research in Plant Cell Compartments* Edited by: Schoefs B. Research Signpost in press.
44. Jomaa H, Wiesner J, Sanderbrand S, Altincicek B, Weidemeyer C, Hintz M, Turbachova I, Eberl M, Zeidler J, Lichtenthaler HK, Soldati D, Beck E: **Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs.** *Science* 1999, **285:**1573-1576.
45. Roos DS, Crawford MJ, Donald RG, Fraunholz M, Harb OS, He CY, Kissinger JC, Shaw MK, Striepen B: **Mining the Plasmodium genome database to define organellar function: what does the apicoplast do?** *Philos Trans R Soc Lond B Biol Sci* 2002, **357:**35-46.
46. Chen F, Mackey AJ, Stoeckert CJ, Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.** *Nucleic Acids Res* 2006, **34:**D363-D368.
47. Bastien O, Ortet P, Roy S, Maréchal E: **A configuration space of homologous proteins conserving mutual information and allo-wing a phylogeny inference based on pair-wise Z-score probabilities.** *BMC Bioinformatics* 2005, **6:**49.
48. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006, **34:**D354-D357.
49. Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P, Krummenacker M, Paley S, Pick J, Rhee SY, Tissier C, Zhang P, Karp PD: **MetaCyc: a multiorganism database of metabolic pathways and enzymes.** *Nucleic Acids Res* 2006, **34:**D511-D516.
50. Ginsburg H: **Progress in in silico functional genomics: the malaria metabolic pathways database.** *Trends Parasitol* 2006, **22:**238-240.
51. **Malaria Parasite Metabolic Pathways** [http://sites.huji.ac.il/malaria/]
52. Kihara D, Skolnick J: **The PDB is a covering set of small protein structures.** *J Mol Biol* 2003, **334:**793-802.
53. **WISDOM, Wide In Silico Docking On Malaria** [http://wisdom.eu-egee.fr/]
54. Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H: **Interpolated Markov models for eukaryotic gene finding.** *Genomics* 1999, **159:**24-31.
55. Pertea M, Salzberg SL, Gardner MJ: **Finding genes in *Plasmodium falciparum*.** *Nature* 2000, **404:**34.
56. Musto H, Rodriguez-Maseda H, Bernardi G: **Compositional properties of nuclear genes from *Plasmodium falciparum*.** *Gene* 1995, **152:**127-132.
57. Musto H, Romero H, Zavala A, Jabbari K, Bernardi G: **Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum* : compositional constraints and translational selection.** *J Mol Evol* 1999, **49:**27-35.
58. Bastien O, Lespinats S, Roy S, Métayer K, Fertil B, Codani JJ, Maréchal E: **Analysis of the compositional biases in *Plasmodium falciparum* genome and proteome using Arabidopsis thaliana as a reference.** *Gene* 2004, **336:**163-173.
59. Karlin S, Altschul SF: **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.** *Proc Natl Acad Sci USA* 1990, **87:**2264-2268.
60. Bastien O, Aude JC, Roy S, Maréchal E: **Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics.** *Bioinformatics* 2004, **20:**534-537.
61. Callebaut I, Prat K, Meurice E, Mornon JP, Tomavo S: **Prediction of the general transcription factors associated with RNA polymerase II in *Plasmodium falciparum* : conserved features and differences relative to other eukaryotes.** *BMC Genomics* 2005, **6:**100.
62. McConkey GA, Pinney JW, Westhead DR, Plueckhahn K, Fitzpatrick TB, Macheroux P, Kappes B: **Annotating the *Plasmodium* genome and the enigma of the shikimate pathway.** *Trends Parasitol* 2004, **20:**60-65.
63. Lipman DJ, Pearson WR: **Rapid and sensitive protein similarity searches.** *Science* 1985, **227:**1435-1441.
64. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147:**195-197.
65. Bacro JN, Comet JP: **Sequence alignment: an approximation law for the *Z-value* with applications to databank scanning.** *Comput Chem* 2001, **25:**401-410.

66. Zhou Y, Young JA, Santrosyan A, Chen K, Yan SF, Winzeler EA: ***In silico* gene function prediction using ontology-based pattern identification.** *Bioinformatics* 2005, **21:**1237-1245.
67. Nagamune K, Sibley LD: **Comparative genomic and phylogenetic analyses of calcium ATPases and calcium-regulated proteins in the apicomplexa.** *Mol Biol Evol* 2006, **23:**1613-1627.
68. Bastien O, Ortet P, Roy S, Maréchal E: **The configuration space of homologous proteins: a theoretical and practical framework to reduce the diversity of the protein sequence space after massive all-by-all sequence comparisons.** *Future Generation Comput Syst* in press.
69. Liu J, Rost B: **Domains, motifs and clusters in the protein universe.** *Curr Opin Chem Biol* 2003, **7:**5-11.
70. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13:**2178-2189.
71. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.
72. Enright AJ, Kunin V, Ouzounis CA: **Protein families and TRIBES in genome sequence space.** *Nucleic Acids Res* 2003, **31:**4632-4638.
73. Yona G, Linial N, Linial M: **ProtoMap: automatic classification of protein sequences and hierarchy of protein families.** *Nucleic Acids Res* 2000, **28:**49-55.
74. Sasson O, Vaaknin A, Fleischer H, Portugaly E, Bilu Y, Linial N, Linial M: **ProtoNet: hierarchical classification of the protein space.** *Nucleic Acids Res* 2003, **31:**348-352.
75. Arnold R, Rattei T, Tischler P, Truong MD, Stumpflen V, Mewes W: **SIMAP – The similarity matrix of proteins.** *Bioinformatics* 2005, **21:**ii42-ii46.
76. Krause A, Stoye J, Vingron M: **Large scale hierarchical clustering of protein sequences.** *BMC Bioinformatics* 2005, **6:**15.
77. Petryszak R, Kretschmann E, Wieser D, Apweiler R: **The predictive power of the CluSTr database.** *Bioinformatics* 2005, **21:**3604-3609.
78. Retief JD: **Phylogenetic analysis using PHYLIP.** *Methods Mol Biol* 2000, **132:**243-258.
79. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18:**502-504.
80. Rastier F: **Ontologie(s).** *Rev Intell Artif* 2004, **18:**15-40.
81. **KEGG: Kyoto Encyclopedia of Genes and Genomes** [http://www.genome.ad.jp/kegg/]
82. Yeh I, Hanekamp T, Tsoka S, Karp PD, Altman RB: **Computational Analysis of *Plasmodium falciparum* Metabolism: Organizing genomic information to facilitate drug discovery.** *Genome Res* 2004, **14:**917-924.
83. **PlasmoCyc** [http://plasmocyc.stanford.edu/]
84. Karp PD, Paley S, Romero P: **The Pathway Tools software.** *Bioinformatics* 2002, **18:**S225-232.
85. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder AA, Batalov S, Carucci DJ, Winzeler EA: **Discovery of gene function by expression profiling of the malaria parasite life cycle.** *Science* 2003, **301:**1503-1508.
86. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*.** *PLoS Biol* 2003, **1:**E5.
87. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder AA, Carucci DJ, Yates JR, Winzeler EA: **Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle.** *Genome Res* 2004, **14:**2308-2318.
88. Le Roch KG, Johnson JR, Ahiboh H, Plouffe D, Henson K, Zhou Y, Ben Mamoun C, Vial H, Winzeler EA: **Genomic profiling of the malaria parasite reponse to the choline analogue reveals drug mechanism of action.** *Proceedings of the Keystone symposia: Malaria: Functional Genomics to Biology to Medicine* 2006.
89. Llinas M, Bozdech Z, Wong ED, Adai AT, DeRisi JL: **Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains.** *Nucleic Acids Res* 2006, **234:**1166-1173.
90. Silvestrini F, Bozdech Z, Lanfrancotti A, Di Giulio E, Bultrini E, Picci L, Derisi JL, Pizzi E, Alano P: **Genome-wide identification of genes upregulated at the onsert of gametocytogenesis in *Plasmodium falciparum*.** *Mol Biochem Parasitol* 2005, **146:**100-110.
91. Ralph SA, Bischoff E, Mattei D, Sismeiro O, Dillies MA, Guigon G, Coppee JY, David PH, Scherf A: **Transcriptome analysis of antigenic variation in Plasmodium falciparum – var gene silencing is not dependent on antisense RNA.** *Genome Biol* 2005, **6:**R93.
92. Daily JP, Le Roch KG, Sarr O, Fang X, Zhou Y, Ndir O, Mboup S, Sultan A, Winzeler EA, Wirth DF: ***In vivo* transcriptional profiling of *Plasmodium falciparum*.** *Malar J* 2004, **3:**30.
93. Daily JP, Le Roch KG, Sarr O, Ndiaye D, Lukens A, Zhou Y, Ndir O, Mboup S, Sultan A, Winzeler EA, Wirth DF: ***In vivo* transcriptome of *Plasmodium falciparum* reveals overexpression of transcripts that encode surface proteins.** *J Infect Dis* 2005, **191:**1196-1203.
94. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney AA, Wolters D, Wu Y, Gardner MJ, Holder AA, Sinden RE, Yates JR, Carucci DJ: **A proteomic view of the *Plasmodium falciparum* life cycle.** *Nature* 2002, **419:**520-526.
95. Fraunholz M: **Systems biology in malaria research.** *Trends Parasitol* 2005, **21:**393-395.
96. Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, Berriman M, Florens L, Janssen CS, Pain A, Christophides GK, James K, Rutherford K, Harris B, Harris D, Churcher C, Quail M, Ormond D, Doggett J, Trueman HE, Mendoza J, Bidwell SL, Rajandream MA, Carucci DJ, Yates JR, Kafatos FC, Janse CJ, Barrell B, Turner CM, Waters AP, Sinden RE: **A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses.** *Science* 2005, **307:**82-86.
97. LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C, Fields S, Hughes RE: **A protein interaction network of the malaria parasite *Plasmodium falciparum*.** *Nature* 2005, **438:**103-107.
98. Suthram S, Sittler T, Ideker T: **The *Plasmodium* protein network diverges from those of other eukaryotes.** *Nature* 2005, **438:**108-112.
99. Wu Y, Wang X, Liu X, Wang Y: **Data-mining approaches reveal hidden families of proteases in the genome of malaria parasites.** *Genome Biol* 2006, **13:**601-616.
100. Young JA, Fivelman QL, Blair PL, de la Vega P, Le Roch KG, Zhou Y, Carucci DJ, Baker DA, Winzeler EA: **The *Plasmodium falciparum* sexual development transcriptome: a microarray analysis using ontology-based pattern identification.** *Mol Biochem Parasitol* 2005, **143:**67-79.
101. **DeRisi Lab Malaria Transcriptome Database** [http://malaria.ucsf.edu]
102. Khachane A, Kumar R, Jain S, Jain S, Banumathy G, Singh V, Nagpal S, Tatu U: **Plasmo2D: an ancillary proteomic tool to aid identification of proteins from *Plasmodium falciparum*.** *J Proteome Res* 2005, **4:**2369-2374.
103. Draghici S, Khatri P, Eklund AC, Szallasi Z: **Reliability and reproducibility issues in DNA microarray measurements.** *Trends Genet* 2006, **22:**101-119.
104. Gibon Y, Usadel B, Blaesing O, Kamlage B, Hoehne M, Trethewey R, Stitt M: **Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes.** *Genome Biol* 2006, **7:**R76.
105. Miron M, Nadon R: **Inferential literacy for experimental high-throughput biology.** *Trends Genet* 2006, **22:**84-89.
106. Shields R: **MIAME, we have a problem.** *Trends Genet* 2006, **22:**65-66.
107. Wang X, Gorlitsky R, Almeida JS: **From XML to RDF: how semantic web technologies will change the design of 'omic' standards.** *Nat Biotechnol* 2005, **23:**1099-1103.
108. Langston MA, Perkins AD, Saxton AM, Scharff JA, Voy BH: **Innovative computational methods for transcriptomic data analysis.** *Proceedings of the ACM Symposium on Applied Computing; Dijon, France* 2006.
109. **European Conferences on Machine Learning and the European Conferences on Principles and Practice of Knowledge Discovery in Databases** [http://lisp.vse.cz/challenge/index.html]
110. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.** *Bioinformatics* 2004, **20:**578-580.
111. Lenhard B, Wahlestedt C, Wasserman WW: **GeneLynx: a gene-centric portal to the human genome.** *Genome Res* 2001, **11:**2151-2157.

112. Beissbarth T, Speed TP: **GOstat: find statistically overrepresented Gene Ontologies within a group of genes.** *Bioinformatics* 2004, **20:**1464-1465.
113. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN: **GoMiner: a resource for biological interpretation of genomic and proteomic data.** *Genome Biol* 2003, **4:**R28.
114. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR: **MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data.** *Genome Biol* 2003, **4:**R7.
115. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4:**P3.
116. Segal E, Yelensky R, Kaushal A, Pham T, Regev A, Koller D, Friedman N: **GeneXPress: A Visualization and Statistical Analysis Tool for Gene Expression and Sequence Data.** *Proceedings of the Eleventh Inter Conf on Intelligent Systems for Molecular Biology* 2004.
117. Lelandais G, Marc P, Vincens P, Jacq C, Vialette S: **MiCoViTo: a tool for gene-centric comparison and visualization of yeast transcriptome states.** *BMC Bioinformatics* 2004, **5:**20.
118. **MADIBA: MicroArray Data Interface for Biological Annotation** [http://www.bi.up.ac.za/MADIBA/]
119. Pir P, Kirdar B, Hayes A, Onsan ZY, Ulgen KO, Oliver SG: **Integrative investigation of metabolic and transcriptomic data.** *BMC Bioinformatics* 2006, **7:**203.
120. Date SV, Stoeckert CJ: **Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale.** *Genome Res* 2006, **16:**542-549.
121. Ohlstein EH, Ruffolo RR Jr, Elliott JD: **Drug discovery in the next millennium.** *Annu Rev Pharmacol Toxicol* 2000, **40:**177-191.
122. Wang S, Sim TB, Kim YS, Chang YT: **Tools for target identification and validation.** *Curr Opin Chem Biol* 2004, **8:**371-377.
123. Freiberg C, Brotz-Oesterhelt H: **Functional genomics in antibacterial drug discovery.** *Drug Discov Today* 2005, **10:**927-935.
124. Freiberg C, Brotz-Oesterhelt H, Labischinski H: **The impact of transcriptome and proteome analyses on antibiotic drug discovery.** *Current Opin Microbiol* 2004, **7:**451-459.
125. Butcher RA, Schreiber SL: **Using genome-wide transcriptional profiling to elucidate small-molecule mechanism.** *Curr Opin Chem Biol* 2005, **9:**25-30.
126. Gunasekera AM, Patankar S, Schug J, Eisen G, Wirth DF: **Drug-induced alterations in gene expression of the asexual blood forms of Plasmodium falciparum.** *Mol Microbiol* 2003, **50:**1229-1239.
127. Birkholtz LM, Claudel-Renard C, Clark K, Louw AI: **Differential transcriptome profiling indicates the physiological significance of polyamines in the human malaria parasite, *Plasmodium falciparum*.** *Proceedings of the Keystone symposia: Malaria: Functional Genomics to Biology to Medicine* 2006.
128. Nirmalan N, Sims PF, Hyde JE: **Quantitative proteomics of the human malaria parasite *Plasmodium falciparum* and its application to studies of development and inhibition.** *Mol Microbiol* 2004, **52:**1187-1199.
129. Makanga M, Bray PG, Horrocks P, Ward SA: **Towards a proteomic definition of CoArtem action in *Plasmodium falciparum* malaria.** *Proteomics* 2005, **5:**1849-1858.
130. Campanale N, Nickel C, Daubenberger CA, Wehlan DA, Gorman JJ, Klonis N, Becker K, Tilley L: **Identification and characterisation of heme-interacting proteins in the malaria parasite, *Plasmodium falciparum*.** *J Biol Chem* 2003, **278:**27354-27361.
131. Knockaert M, Gray N, Damiens E, Chang YT, Grellier P, Grant K, Fergusson D, Mottram J, Soete M, Dubremetz JF, Le Roch K, Doerig C, Schultz P, Meijer L: **Intracellular targets of cyclin-dependent kinase inhibitors: identification of affinity chromatography using immobilized inhibitors.** *Chem Biol* 2000, **7:**411-422.
132. Graves PR, Kwiek JJ, Fadden P, Ray R, Hardeman K, Coley AM, Foley M, Haystead TA: **Discovery of novel targets of quinoline drugs in the human purine binding proteome.** *Mol Pharmacol* 2002, **62:**1364-1372.
133. Oliver S: **Guilt-by-association goes global.** *Nature* 2000, **403:**601-603.
134. Voy BH, Scharff JA, Perkins AD, Saxton AM, Borate B, Chesler EJ, Branstetter LK, Langston MA: **Extracting gene networks for low-dose radiation using graph theoretical algorithms.** *PLoS Comput Biol* 2006, **21(2):**e89.
135. Zhou Z, Schnake P, Xiao L, Lal AA: **Enhanced expression of a recombinant malaria candidate vaccine in *Escherichia coli* by codon optimization.** *Protein Expr Purif* 2004, **34:**87-94.
136. Llinas M, del Portillo HA: **Mining the malaria transcriptome.** *Trends Parasitol* 2005, **21:**350-352.
137. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL: **Global mapping of pharmacological space.** *Nat Biotechnol* 2006, **24:**805-815.
138. Sugiyama T, Suzue K, Okamoto M, Inselburg J, Tai K, Horii T: **Production of recombinant SERA proteins of *Plasmodium falciparum* in *Escherichia coli* by using synthetic genes.** *Vaccine* 1996, **14:**1069-1076.
139. Withers-Martinez C, Carpenter EP, Hackett F, Ely B, Sajid M, Grainger M, Blackman MJ: **PCR-based gene synthesis as an efficient approach for expression of the A+T-rich malaria genome.** *Protein Eng* 1999, **12:**1113-1120.
140. Yadava A, Ockenhouse CF: **Effect of codon optimization on expression levels of a functionally folded malaria vaccine candidate in prokaryotic and eukaryotic expression systems.** *Infect Immun* 2003, **71:**4961-4969.
141. Flick K, Ahuja S, Chene A, Bejarano MT, Chen Q: **Optimized expression of *Plasmodium falciparum* erythrocyte membrane protein 1 domains in *Escherichia coli*.** *Malar J* 2004, **3:**50.
142. Christopherson RI, Cinquin O, Shojaei M, Kuehn D, Menz RI: **Cloning and expression of malarial pyrimidine enzymes.** *Nucleosides Nucleotides Nucleic Acids* 2004, **23:**1459-1465.
143. Mehlin C, Boni E, Buckner FS, Engel L, Feist T, Gelb MH, Haji L, Kim D, Liu C, Mueller N, Myler PJ, Reddy JT, Sampson JN, Subramanian E, Van Voorhis WC, Worthey E, Zucker F, Hol WG: **Heterologous expression of proteins from *Plasmodium falciparum* : Results from 1000 genes.** *Mol Biochem Parasitol* 2006, **148:**144-160.
144. Kihara D, Skolnick J: **The PDB is a covering set of small protein structures.** *J Mol Biol* 2003, **334:**793-802.
145. **SGC: Structural genomics consortium** [http://www.sgc.utoronto.ca/]
146. **SGPP: Structural Genomics of Pathogenic Protozoa** [http://www.sgpp.org/]
147. Birkholtz LM, Wrenger C, Joubert F, Wells GA, Walter RD, Louw AI: **Parasite-specific inserts in the bifunctional S-adenosylmethionine decarboxylase/ornithine decarboxylase of *Plasmodium falciparum* modulate catalytic activities and domain interactions.** *Biochem J* 2004, **377:**439-448.
148. Wells GA, Birkholtz LM, Joubert F, Walter RD, Louw AI: **Novel properties of malarial S-adenosylmethionine decarboxylase as revealed by structural modelling.** *J Mol Graph Model* 2006, **24:**307-318.
149. **MEME: Multiple Em for Motif Elicitation** [http://meme.sdsc.edu/meme/]
150. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs of biopolymers.** In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* AAAI Press; 1994:28-36.
151. de Beer TA, Louw AI, Joubert F: **Elucidation of sulfadoxine resistance with structural models of the bifunctional *Plasmodium falciparum* dihydropterin pyrophosphokinase-dihydropteroate synthase.** *Bioorg Med Chem* 2006, **14:**4433-4443.
152. Bastien O, Roy S, Maréchal E: **Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions.** *C R Biol* 2005, **328:**445-453.
153. Toyoda T, Reynolds KBB, Gen-ichiro S, Toshihiro H, Nobuo T, Akiko I: **Lead discovery of inhibitors of the dihydrofolate reductase domain of *Plasmodium falciparum* dihydrofolate reductase-thymidylate synthase.** *Biochem Biophys Res Commun* 1997, **235:**515-519.
154. McKie JH, Douglas KT, Chan C, Roser SA, Yates R, Read M, Hyde JE, Dascombe MJ, Yuthavong Y, Sirawaraporn W: **Rational drug design approach for overcoming drug resistance: application to pyrimethamine resistance in malaria.** *J Med Chem* 1998, **41:**1367-1370.
155. Lemcke T, Christensen IT, Jøgensen FS: **Towards and understanding of drug resistance in malaria: three-dimensional structure of *Plasmodium falciparum* dihydrofolate reductase by homology building.** *Bioorg Med Chem* 1999, **7:**1003-1011.

156. Rastelli G, Sirawaraporn W, Sompornpisut P, Vilaivan T, Kamchonwongpaisan S, Quarrell R, Lowe G, Thebtaranonth Y, Yuthavong Y: **Interaction of pyrimethamine, cycloguanil, WR99210 and their analogues with *Plasmodium falciparum* dihydrofolate reductase: structural basis of antifolate resistance.** *Bioorg Med Chem* 2000, **8:**1117-1128.
157. Santos-Filho OA, de Alencastro RB, Figueroa-Villar JD: **Homology modeling of wild type and pyrimethamine/cycloguanil-cross resistant mutant type *Plasmodium falciparum* dihydrofolate reductase A model for antimalarial chemotherapy resistance.** *Biophys Chem* 2001, **91:**305-317.
158. Delfino TR, Santos-Filho OA, Figueroa-Villar JD: **Molecular modeling of wild-type and antifolate resistant mutant *Plasmodium falciparum* DHFR.** *Biophys Chem* 2002, **98:**287-300.
159. Yuvaniyama J, Citnumsub P, Kamchonwongpaisan S, Vanichtanankul J, Sirawaraporn W, Taylor P, Walkinshaw MD, Yuthavong Y: **Insights into antifolate resistance from malarial DHFR-TS structures.** *Nature* 2003, **10:**357-65.
160. Li R, Chen X, Gong B, Selzer PM, Li Z, Davidson E, Kurzban G, Miller RE, Nuzum EO, McKerrow JH, Fletterick RJ, Gillmor SA, Craik CS, Kuntz ID, Cohen FE, Kenyon GL: **Structure-based design of parasitic protease inhibitors.** *Bioorg Med Chem* 1996, **4:**1421-1427.
161. Desai PV, Patny A, Sabnis Y, Tekwani B, Gut J, Rosenthal P, Srivastava A, Avery M: **Identification of novel parastic cysteine protease inhibitors using virtual screening. 1. The Chembridge database.** *J Med Chem* 2004, **47:**6609-6615.
162. Desai PV, Patny A, Gut J, Rosenthal PJ, Tekwani B, Srivastava A, Avery M: **Identification of novel parasitic cysteine protease inhibitors by use of virtual screening. 2. The Available Chemical Directory.** *J Med Chem* 2006, **49:**1576-1584.
163. Gutiérrez-de-Terán H, Nervall M, Ersmark K, Liu P, Janka LK, Dunn B, Hallberg A, Åqvist J: **Inhibitor binding to the Plasmepsin IV aspartic protease from *Plasmodium falciparum*.** *Biochemistry* 2006, **45:**10529-10541.
164. **ICGEB: The International Center for Genetic Engineering and Biotechnology** [http://net.icgeb.org/]
165. Kellenberger E, Muller P, Schalon C, Bret G, Foata N, Rognan D: **sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank.** *J Chem Inf Model* 2006, **46:**717-727.
166. Russ AP, Lampel S: **The druggable genome: an update.** *Drug Discov Today* 2005, **10:**1607-1610.
167. Hajduk PJ, Huth JR, Tse C: **Predicting protein druggability.** *Drug Discov Today* 2005, **10:**1675-1682.
168. Doolan DL, Southwood S, Freilich DA, Sidney J, Graber NL, Shatney L, Bebris L, Florens L, Dobano C, Witney AA, Appella E, Hoffman SL, Yates JR, Carucci DJ, Sette A: **Identification of *Plasmodium falciparum* antigens by antigenic analysis of genomic and proteomic data.** *Proc Natl Acad Sci USA* 2003, **100:**9952-9957.
169. Doolan DL, Aguiar JC, Weiss WR, Sette A, Felgner PL, Regis DP, Quinones-Casas P, Yates JR, Blair PL, Richie TL, Hoffman SL, Carucci DJ: **Utilization of genomic sequence information to develop malaria vaccines.** *J Exp Biol* 2003, **206:**3789-3802.
170. Buyya R, Branson K, Giddy J, Abramson D: **The Virtual Laboratory. A Toolset to Enable Distributed Molecular Modeling for Drug Design on the WorldWide Grid.** *Concurrency Computat: Pract Exper* 2003, **15:**1-25.
171. Chien A, Foster I, Goddette D: **Grid technologies empowering drug discovery.** *Drug Discov Today* 2002, **7(Suppl 20):**176-180.
172. Garcia-Aristegui DJ, Mendez Lorenzo P, Valverde JR: **GROCK: High-Throughput Docking Using LCG Grid Tools.** *Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing* 2005:85-90.
173. Sudholt W, Baldridge KK, Abramson D, Enticott C, Garic S, Kondric C, Nguyen D: **Application of grid computing to parameter sweeps and optimizations in molecular modelling.** *Future Generation Comput Syst* 2005, **21:**27-35.
174. Peitsch MC, Morris GE, Basse-Welker J, Cartwright G, Juterbock D, Marti KO, Lorban S, Odell G, Vachon T: **Informatics and Knowledge Management at the Novartis Institutes for BioMedical Research.** *SCIP-online* 2004, **46:**1-4.
175. Richards WG: **Virtual screening using grid computing: the screensaver project.** *Nat Rev Drug Discov* 2002, **1:**551-555.
176. **EGEE: Enabling Grid for E-sciencE** [http://public.eu-egee.org/]
177. Francis SE, Sullivan DJ Jr, Goldberg DE: **Hemoglobin metabolism in the malaria parasite *Plasmodium falciparum*.** *Annu Rev Microbiol* 1997, **51:**97-123.
178. Coombs GH, Goldberg DE, Klemba M, Berry C, Kay J, Mottram JC: **Aspartic proteases of *Plasmodium falciparum* and other protozoa as drug targets.** *Trends Parasitol* 2001, **17:**532-537.
179. Silva AM, Lee AY, Gulnik SV, Majer P, Collins J, Bhat TN, Collins PJ, Cachau RE, Luker KE, Gluzman IY, Francis SE, Oksman A, Goldberg DE, Erickson JW: **Structure and inhibition of plasmepsin II, A haemoglobin degrading enzyme from *Plasmodium falciparum*.** *Proc Natl Acad Sci USA* 1996, **93:**10034-10039.
180. Jiang S, Prigge ST, Wei L, Gao Y, Hudson TH, Gerena L, Dame JB, Kyle DE: **New class of small nonpeptidyl compounds blocks *Plasmodium falciparum* development *in vitro* by inhibiting plasmepsins.** *Antimicrob Agents Chemother* 2001, **45:**2577-2584.
181. Lipinski CA: **Chris Lipinski discusses life and chemistry after the Rule of Five.** *Drug Discov Today* 2003, **8:**12-16.
182. **ACCAMBA** [http://accamba.imag.fr/]
183. Feldman HJ, Dumontier M, Ling S, Haider N, Hogue CW: **CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules.** *FEBS Lett* 2005, **579:**4685-4691.
184. **ChEBI: Chemical Entities of Biological Interest** [http://www.ebi.ac.uk/chebi/]
185. **The PubChem Project** [http://pubchem.ncbi.nlm.nih.gov/]
186. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2005, **33:**D154-D159.
187. **The UniProt Knowledge Base** [http://www.expasy.uniprot.org/]
188. Ofran Y, Punta M, Schneider R, Rost B: **Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery.** *Drug Discov Today* 2005, **10:**1475-1482.
189. Breton V, Jacq N, Hofmann M: **Grid added value to address malaria.** *Proceedings of Biogrid Workshop, CCGRID conference: May 2006; Singapore* 2006.
190. **@neuRIST: Integrated Biomedical Informatics for the Management of Cerebral Aneurysms** [http://www.aneurist.org]
191. Wood V, Rutherford KM, Ivens A, Rajandream MA, Barrell B: **A Re-annotation of the *Saccharomyces cerevisiae* genome.** *Comp Funct Genom* 2001, **2:**143-154.
192. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408:**796-815.
193. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2001, **431:**931-945.
194. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V,

Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.

195. **Ensembl genome browser** [http://www.ensembl.org/index.html]
196. **PlasmoDB: Malaria Parasite Genome Project** [http://www.plasmodb.org/plasmo/home.jsp]