# Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale

Shailesh V. Date[1] and Christian J. Stoeckert Jr.

*Center for Bioinformatics, Department of Genetics, School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*

**Many thousands of proteins encoded by the genome of *Plasmodium falciparum*, the causal organism of the deadliest form of human malaria, are of unknown function. It is of utmost importance that these proteins be characterized if we are to develop combative strategies against malaria based on the biology of the parasite. In an attempt to infer protein function on a genome-wide scale, we computationally modeled the *P. falciparum* interactome, elucidating local and global functional relationships between gene products. The resulting interaction network, reconstructed by integrating in silico and experimental functional genomics data within a Bayesian framework, covers ~68% of the parasite genome and provides functional inferences for more than 2000 uncharacterized proteins, based on their associations. Network reconstruction involved the use of a novel strategy, where we incorporated continuously updated, uniform reference priors in our Bayesian model. This method for generating interaction maps is thus also well suited for application to other genomes, where pre-existing interactome knowledge is sparse. Additionally, we superimposed this map on genomes of three apicomplexan pathogens—*Plasmodium yoelii*, *Toxoplasma gondii*, and *Cryptosporidium parvum*—describing relationships between these organisms based on retained functional linkages. This comparison provided a glimpse of the highly evolved nature of *P. falciparum*; for instance, a deficit of nearly 26% in terms of predicted interactions is observed against *P. yoelii*, because of missing ortholog partners in pairs of functionally linked proteins.**

**[Supplemental material is available online at www.genome.org and results from this study are available for download from http://cbil.upenn.edu/plasmoMAP/.]**

The genome sequence of *Plasmodium falciparum*, the causative organism of the deadliest form of human malaria, has revealed many surprising details about the parasite, including the novel nature of its many genes. More than 60% of the genome is as yet uncharacterized; a majority of the genes bear no acceptable sequence homology with known genes in other organisms (Gardner et al. 2002). If we are to develop effective control strategies against malaria based on parasite biology, it is essential that we characterize these many unknown genes and their products and understand the interactions between them, both locally and on a genome-wide scale.

Here we describe computational modeling of the *Plasmodium falciparum* interactome, which reveals local and genome-wide functional relationships between proteins in the parasite genome and permits functional assignments based on associations between characterized and uncharacterized proteins. The interactome, as captured by a network of pairwise functional linkages, was reconstructed by integrating data from publicly available *P. falciparum* transcriptome profiling studies and linkages generated in silico within a Bayesian framework. Data integration for inferring protein associations proves advantageous for two well-known reasons—first, combining data from diverse, large-scale studies of genome, which vary considerably in their accuracy when compared against standardized sets, generates data sets of higher quality (von Mering et al. 2002; Jansen et al. 2003; Lee et al. 2004), and second, data integration effectively captures different aspects of the parasite biology, as reported by the individual methods (Lee et al. 2004). The *P. falciparum* interaction network thus contains a proportion of functional links that are supported by multiple lines of evidence and includes functional relationships arising from direct physical contact, such as membership in the same physical complex, as well as more subtle relationships, such as linkages due to membership in the same cellular pathway or system.

The Bayesian framework we employ yields a likelihood score for individual protein pairs as a measure of goodness. Given reliable standards of reference, often referred to as "gold standards," and reasonably estimated prior odds, such naive Bayesian models and their variants have recently been used to successfully integrate diverse data types, yielding probabilistic measurements of yeast interactome (Jansen et al. 2003; Troyanskaya et al. 2003; Lee et al. 2004). Data integration is primarily achieved by measuring individual accuracies in the context of overlap with the gold standards, then substituting or supplementing them with appropriate weights or scores. This simultaneously allows for standardization and incorporation of expert knowledge in the model. The likelihood of a pair of genes being functionally linked, within this scheme, is thus an estimate of the combined strength of linkages for the pair from all included data sets. This method for network reconstruction, outlined in Figure 1A, reveals a general strategy that is highly suitable for application to other minimally studied genomes as well, where protein–protein interaction data is sparse. The reconstructed *P. falciparum* interaction network provides functional information for 3667 genes or their products, at varying levels of confidence. We describe

[1]**Corresponding author.**
**E-mail svdate@pcbi.upenn.edu; fax (215) 573-3111.**
Article published online ahead of print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.4573206. Freely available online through the *Genome Research* Open Access option.


16:542–549 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

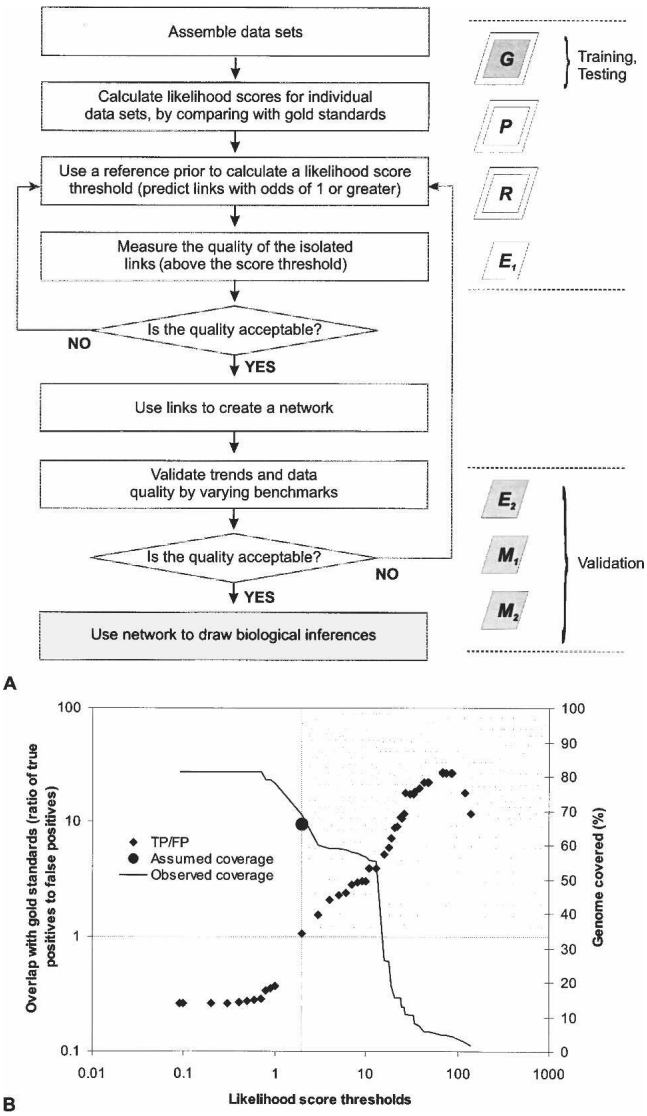


**Figure 1.** A generalized strategy for integrating diverse functional genomics data and reconstructing functional interaction networks. (*A*) A salient feature of this strategy involves the use of reference priors containing an assumed number of links, in the absence of knowledge about gene interactions in a genome. An appropriate prior that accurately reflects the state of genome can then be used to isolate a confident result set. (G) Benchmark gold standards; (P) phylogenetic profile data; (R) Rosetta stone linkage data; (E) gene expression data ($E_1$: Bozdech et al. 2003; $E_2$: Le Roch et al. 2003); (M) mass spectrometric data ($M_1$ and $M_2$). Dark boxes indicate benchmarks; boxes with double borders indicate in silico data. (*B*) Prior knowledge about the possible number of functionally interacting proteins leads to a corresponding likelihood score threshold, with posterior odds of finding a true pair of functionally linked proteins greater than one. Likelihood score thresholds and the ratio of true to false positives were calculated as described by Jansen et al. (2003). For our chosen threshold, the number of proteins assumed to functionally interact with each other corresponds very closely with the number of proteins with predicted functional linkages in our final data set. Linkages from the shaded area were used for network reconstruction. (Cross-hair) Chosen likelihood score threshold and estimated accuracy; (●) approximate coverage described by our prior belief; (♦) ratio of true positives to false positives.

several examples of how the map can be exploited to draw meaningful biological conclusions, and how, in one case, it validates the role of RESA proteins in heat-shock response. These interactions represent a knowledge bank, which can be used to elucidate the biology of the parasite and design experiments aimed at gene characterization. All results of this study are available at http://cbil.upenn.edu/plasmoMAP/ and will also be made available through the PlasmoDB Web site (http://www.plasmodb.org; Kissinger et al. 2002) as predicted linkages for individual genes.

## Results

### Naive Bayesian integration of in silico and functional genomics data

Amino acid sequences of 5334 known or predicted *P. falciparum* proteins were compared against 163 completely sequenced eukaryotic and prokaryotic genomes, using BLAST. These data were used to construct phylogenetic profiles (Pellegrini et al. 1999) and mined along with BLAST data from one additional genome for the presence of Rosetta stone fusion proteins (Marcotte et al. 1999). After appropriate filtering, we retained phylogenetic profiles and Rosetta stone fusion information for 2813 and 993 proteins, respectively (see Supplemental information). We also calculated pairwise correlations between 3471 normalized gene expression profiles generated by large-scale microarray analysis of the intra-erythrocytic developmental cycle (IDC) transcriptome of *P. falciparum*, published by others (Bozdech et al. 2003, HB3 strain). Of all available microarray expression data sets, this set affords the highest resolution of the parasite transcriptome over 48 time points and therefore yields more accurate correlation values, although it does not include information about other life cycle stages. Together, these three data sets capture functional information for 4343 unique proteins (~81% of the proteome) from all life-cycle stages of the parasite. A set of 655 proteins is contained in all functional genomics data sets, whereas 2279 proteins are represented in at least two individual sets.

Positive and negative gold standards were derived using pathway assignments from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al. 2004), combined with data from Gene Ontology Process assignments (GO process) (The Gene Ontology Consortium 2004) provided by the sequencing centers. We obtained 10,267 positive gold standard pairs ($G_P$) and 44,812 negative gold standard pairs ($G'_N$) from these sources, containing information from at least 57 distinct KEGG pathways, based on Enzyme commission (EC) number annotation of 2 or more components from each pathway (see Methods and Supplemental information). Likelihood scores for individual functional genomics data sets were calculated by measuring the overlap between binned confidence measurements from each data set and the gold standards. Final likelihood scores for each pair were a product of likelihood scores from all individual data sets where the particular pair was observed. Functional linkage data for two other *P. falciparum* strains, 3D7 and Dd2, are provided at http://cbil.upenn.edu/plasmoMAP/ (see also Supplemental information).

### Estimating prior odds and prediction accuracy

Overall, we calculated likelihood scores for 8,019,065 pairwise functional linkages, where evidence from one or more functional genomics data sets was available. For this set of predicted links, choosing a likelihood score threshold for accepting results depended on accurately estimating the number of functionally interacting proteins, representing our prior odds. Given the largely unknown nature of the *P. falciparum* genome, this number is difficult to determine. We therefore used a series of estimates of

the number of functional interactions likely in parasite genome, as our reference priors. These estimates of evidence were used to derive likelihood score thresholds (see Methods for the relationship between likelihood ratios and prior odds; see also Supplemental information). For each reference prior, we predicted functional linkages with posterior odds of one or greater and calculated the accuracy of the result set by measuring the ratio of observed true and false positives, based on comparison with the gold standards. This procedure allowed us to accommodate all prior beliefs regarding the possible number of links, estimate accuracy of each set, and choose a relevant set of functional links for map reconstruction. For the *P. falciparum* genome, these assumptions are relevant for a broad numerical range of possible linkages; however, they might prove ineffective if prior beliefs hold that a very large or very small percentage of proteins functionally interact with each other.

We accepted all links above a likelihood score threshold of two, below which the ratio of true to false positives was less than one, corresponding to the belief that approximately two-thirds of all *P. falciparum* proteins are a part of the interactome (Fig. 1B). This generated 388,969 functional linkages between 3667 proteins, permitting description of ~68% of the *P. falciparum* genome at varying levels of confidence. At least 69 KEGG pathways are represented in this data set, based on EC number annotation of two or more components, and ~30% (117,764) of all interactions are supported by two or more lines of evidence (Table 1). Initial quality of the functional linkages was determined by comparing predicted links with different sets of randomized protein pairs (see Methods). Taken together, these tests reveal a high average accuracy of ~93% associated with the predicted functional linkages (Fig. 2A). As more rigorous tests, we formally verified our results using sevenfold cross-validation and gain of information using normal and shuffled sets of input data. The results of these tests, as illustrated in Figure 2B, indicate a very large gain of information over the shuffled input set, further attesting the high quality of the integrated data (see Methods and Supplemental information for details).

### Confirming network robustness to changing benchmarks

To assess the robustness of our predictions to changing benchmarks and to detect overtraining, we compared information in the predicted network with extraneous gene expression data (Le Roch et al. 2003) and two independent mass spectrometric measurements of the proteome (Florens et al. 2002; Lasonder et al. 2002; hereafter referred to as sets $M_1$ and $M_2$, respectively). Specifically, we first polled the three large-scale data sets for individual proteins that do not share the same life-cycle stages, that is, proteins detected in only one of the sporozoite, gamete/gametocyte, or erythrocyte-associated life-cycle stages where available (we collectively refer to proteins not found in the sporozoite or gamete/gametocyte life-cycle stages as proteins in erythrocyte-associated stages for convenience). We then paired proteins from similar and dissimilar life-cycle stages to create additional benchmarks of positive and negative sets, respectively, and compared these with functional linkages included in our predicted set.

As a part of our input data was restricted to proteins in the erythrocyte-associated stages, we first examined the interaction network for biases with respect to representation of linkages from various life-cycle stages of the parasite. For this analysis, we included information about genes expressed only in the sporozoite and gamete/gametocytic stages, based on data published by Le Roch et al. (2003), and measured the agreement between the positive and negative sets and our predicted network. Likelihood scores show excellent agreement with these data; the number of positives represented in our predictions increase corresponding to likelihood scores, showing a general trend similar to the integrated data (Fig. 2C). The map therefore substantially represents functional relationships between proteins from all life-cycle stages, without any bias. A similar exercise with the $M_1$ and $M_2$ data sets reveals an increase in the number of predictions that overlap with the positive set, underscoring the fact that pairs with increasing likelihood scores gain accuracy, and the model is robust to a variation of benchmarks (Fig. 2D,E). Differences observed when dealing with the $M_1$ and $M_2$ data sets and results of a test for spatial separation of linkages based on life-cycle stages are discussed in Supplemental information.

**Table 1. Attributes of selected data sets**

| | Interaction map | High-confidence subset |
|---|---|---|
| Number of unique pairs | 388,969 | 12,290 |
| Number of unique proteins | 3667 | 1415 |
| Minimum true positive to false positive ratio | 1 | 10 |
| Number of unknowns | 2216 | 507 |
| unknowns linked to knowns | 2109 | 410 |
| unknowns linked to unknowns | 107 | 97 |
| Number of pathways represented (with more than two components)[a] | 69 | 56 |
| Lines of evidence (by pairs) | | |
| from three sets | 1148 (<1%) | 502 (~4%) |
| from two or more sets | 117,764 (~30%) | 12,034 (~97%) |
| from one set | 271,205 (~70%) | 256 (~2%) |

Pairs included in the interaction map were selected based on the total likelihood scores that were ≥2, while pairs in the subset of high-confidence links were selected for likelihood scores ≥14. Details of information about pathways represented in these sets are available at http://cbil.upenn.edu/plasmoMAP/.
[a]The number of shared pathways is described based on EC number annotation of KEGG pathway components.

## Discussion

Several models based on Bayesian integration of data have been developed and used to describe interactome networks in well-studied genomes such as yeast. These have benefited from the availability of high-quality curated interaction data (e.g., see the Database of Interacting Proteins; Salwinski et al. 2004). In the absence of such interaction data, we have used KEGG database and GO process annotations to create a set of gold standards, the effectiveness of which has been demonstrated previously (Pellegrini et al. 1999; Date and Marcotte 2003). One possible drawback of using KEGG pathways is the introduction of a systemic bias in favor of more metabolic knowledge in training and testing. Although we cannot discount such a bias, it is difficult to estimate its extent in our predictions. Other errors, when directly using linkages included in this map, might arise from our use of expression data from a single life-cycle stage and the use of simple measures like Pearson correlation to calculate similarity between expression profiles. As more functional genomics

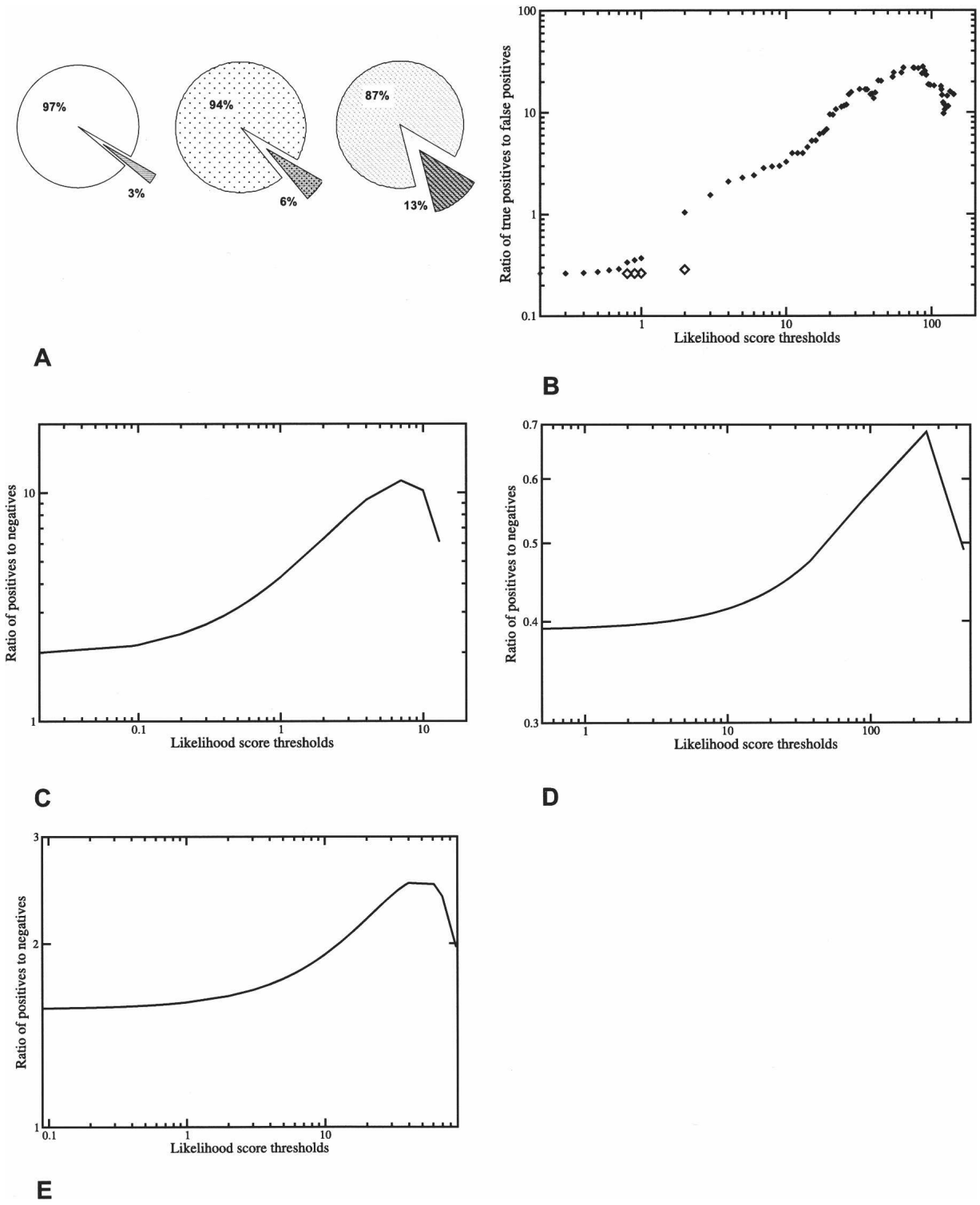


**Figure 2.** Quality of the predicted set of functional linkages and comparison with other functional genomics data sets. (*A*) Initial data quality was assessed using sets of randomly paired proteins (see Methods). Each graph indicates overlap between random sets and predicted linkages; darker shades indicate false positives or overlap with random sets; (○) all possible *P. falciparum* pairs; (⊙) proteins in network; (◎) shuffled pairs from the map. (*B*) Tests using sevenfold cross-validation and comparison with likelihood scores from shuffled input data indicate the suitability of the predicted functional linkages for inclusion in the interaction map. Likelihood scores from the shuffled sets, indicated as (◇), are significantly lower, and show a very high percentage of false positives, attesting to the information gain when using normal linkages. The model is robust to changing benchmarks. Similar trends are generated when tested with data generated by Le Roch et al. (2003) (tested with 21,486 positives and 8160 negatives) (*C*), the $M_1$ (tested with 267,716 positives and 699,029 negatives) (*D*), and $M_2$ (tested with 191,376 positives and 129,825 negatives) (*E*) data sets. A line (*C,D,E*) represents an analytical fit to the data by least squares.

data becomes available, however, the model can be retrained to take advantage of new information, in turn making the predictions more robust. With proper testing, advanced measures of profile comparison, such as mutual information, could also be applied to the expression data. Meanwhile, it is advisable to interpret functional relationships in the light of other kinds of biological knowledge, including, for instance, chromosomal location, to avoid generating conflicting biological scenarios.

## Exploring the *P. falciparum* interactome

The degree of coverage and sensitivity of ~21% are comparable to results discussed in recently published studies of the yeast genome that use Bayesian integration schemas (Jansen et al. 2003; Troyanskaya et al. 2003; Lee et al. 2004). Overall, ~60% of all proteins in the reconstructed interaction map are currently annotated as hypothetical, of which about 95% (2109/2216) are seen linked to other known proteins. One hundred seven hypothetical proteins are linked only with other hypothetical proteins, potentially representing new pathways or previously uncharacterized components of known biochemical pathways. An interesting example involves functional links between PF14_0123, an uncharacterized protein, and members of the TCP-1 chaperonin-containing proteins (Fig. 3A). Members of the TCP-1 ring complex (also referred to as TRiC or CCT) are group II chaperonins conserved in Archaea and Eukarya and contain two ring assemblies that help fold proteins that cannot be folded by simpler chaperone systems. Besides a link to PF14_0574, another uncharacterized protein, five of the six proteins to which PF14_0123 is linked represent different subunits of the TCP-1 complex, strongly suggesting an association with the TCP-1 complex proteins and a role in the protein-folding machinery.

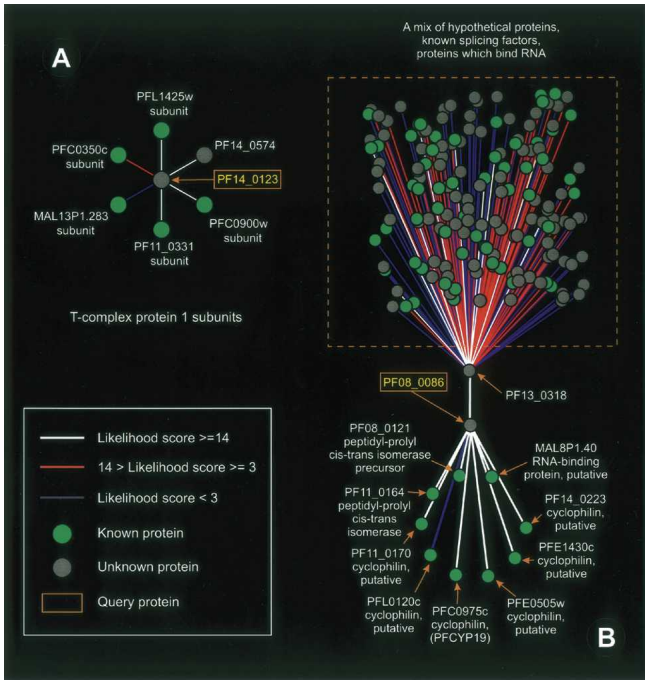


**Figure 3.** Network neighborhoods of PF14_0123 and PF08_0086. Associations with known proteins suggest a strong role for PF14_0123 (*A*) in the protein-folding machinery, whereas PF08_0086 (*B*) is likely involved in protein folding and trafficking. PF08_0086 is also linked to an uncharacterized protein associated with splicing factors and RNA-binding proteins. Together, PF08_0086 and PF13_0138 are likely a part of a larger cellular system involving protein synthesis and transport. Links were visualized using Cytoscape (Shannon et al. 2003). (Link colors) Phase transitions; (White links) highest confidence subset (see Results and Fig. 4).

PF08_0086, another uncharacterized protein, is mostly seen linked to cyclophilins (Fig. 3B). Of all proteins in this set of links, six are annotated as cyclophilins, while two are annotated as precursors to peptidyl–prolyl *cis–trans* isomerases. Cyclophilins are known *cis/trans* isomerases that isomerize peptidyl–prolyl bonds in peptides (Galat and Metcalfe 1995). These links suggest a role for PF08_0086 in protein folding and trafficking. Further, PF08_0086 is also linked to PF13_0318, another uncharacterized protein, which in turn is functionally linked to known splicing factors and RNA-binding proteins, suggesting a role for PF13_0318 in transcription, mRNA processing, and translation initiation. Taken together, functional associations between these proteins reveal components that are actively engaged in protein synthesis and transport and probably are parts of a larger cellular system. The value of these associations is further enhanced given the fact that cyclophilins have been considered possible chemotherapeutic targets for treatment of malaria (Gavigan et al. 2003); such analyses will undoubtedly prove useful in further elucidating the downstream mechanism of action of cyclosporins and other immunosuppressive drugs.

Linkage data for both uncharacterized proteins PF14_0123 and PF08_0086 are supported by evidence from two different functional genomics data sets. Similar analyses for other unknown proteins reveal functional linkages that can be used for function assignment (see Supplemental information for more examples).

## Analyzing a subset of higher confidence links

When the distribution of proteins and protein–protein linkages were examined with regard to the likelihood score thresholds, we noticed phase transitions within the data, where the number of functional interactions at a particular threshold increased by a factor of two or more. We exploited these transitions to isolate and describe a subset of our data with higher confidence than all data combined. This data set includes functional links above a likelihood score threshold of 14 or greater, representing 12,290 links between 1415 unique proteins, and two or more components from 56 different KEGG pathways. Linkages above this threshold include 10× more true positives than false positives, by the test of Figure 1. When examined for the presence of false positives generated by randomly pairing proteins similar to our previous tests, nearly 96% of such pairs were excluded from this set, as were ~95% of the shuffled pairs. This indicates the very high accuracy of the included functional linkages, which also represent the core of our interaction network.

Visualization of these links (Adai et al. 2004, see Supplemental information) reveals components of various systems, along with clusters of mixed proteins (Fig. 4). A group of proteins that predominantly consists of DnaJ domain proteins is also visible. Based on this grouping, hypothetical proteins PFB0090c, PF14_0359, MAL6P1.16, PF11_0026, PF10_0388, PF07_0002, PFC1075w, and PFB0980w, almost all containing only a weak DnaJ motif, appear to be putative members of the heat-shock response machinery. Interestingly, a few proteins in the cluster also contain RESA-like domains. Past studies have speculated on the role of the RESA family given the prominent presence of the DnaJ domain in these proteins (Watanabe 1997). Although it is

possible that in our analyses we observe a grouping of proteins involved in the heat-shock machinery only because of the common presence of the DnaJ domain, the grouping does suggest that members of the RESA family of proteins may play some role in the parasite response to heat shock. A recent study confirms these speculations; Silva et al. (2005) report pronounced susceptibility of *resa1* knock-out parasites to heat shock, supporting the conclusions drawn based on observed functional interactions in our data set. Proximity to regions rich in the antigenically important var and Rifin proteins might also indicate a putative role of the RESA family in antigenic variation (Marti et al. 2004). Examples of insights into the functions of other proteins are discussed in Supplemental information, along with details of the transitions.

### Linkage detection in other apicomplexa

In addition to interpreting functional associations in *P. falciparum*, we also examined retention of linkages in genomes of three other apicomplexan species: *P. yoelii*, *Toxoplasma gondii*, and *Cryptosporidium parvum*. Superimposition of the interaction map on these genomes reveals commonalities and differences between the organisms; by identifying linkages between a conserved pair of orthologs, we are able to ascertain relationships between organisms based not just on the presence or absence of a gene or changes in sequence but more so in terms of an evolutionary pressure for link retention.

Ortholog substitutions (OrthoMCL, Li et al. 2003; see also OrthoMCL-DB, Chen et al. 2006) reveal that nearly 12% of *P. falciparum* functional linkages are retained across all four genomes, while 87,798 linkages represented by 3307 proteins (see Table 2), are exclusive to *P. falciparum*. Interesting biological observations based on these analyses are discussed in Supplemental information. Intuitively, *P. falciparum* and *P. yoelli* are expected to share the most number of links, as is evident from the data. Interestingly, however, the same intuition underestimates the exact significance of the differences between these two species. A deficit of nearly 26% in terms of the predicted interactions exists between *P. falciparum* and *P. yoelii*, where an ortholog for one partner in a pair of functionally linked proteins was absent, further highlighting the well-known differences between *P. falciparum* and *P. yoelii* (Carlton et al. 2002). It is likely that a majority of these differences are due to the presence of the var and rif/stevor families significantly absent in the *P. yoelii* genome, along with genes in other less-represented categories, such as the various physiological process required for survival in a specialized habitat, cell–cell communication, and even aspects of invasion and adhesion (Carlton et al. 2002). Examination of these unique linkages reveals mostly functional associations involving hypothetical proteins, suggesting that the many unknown proteins in this organism play a significant role in its biology.

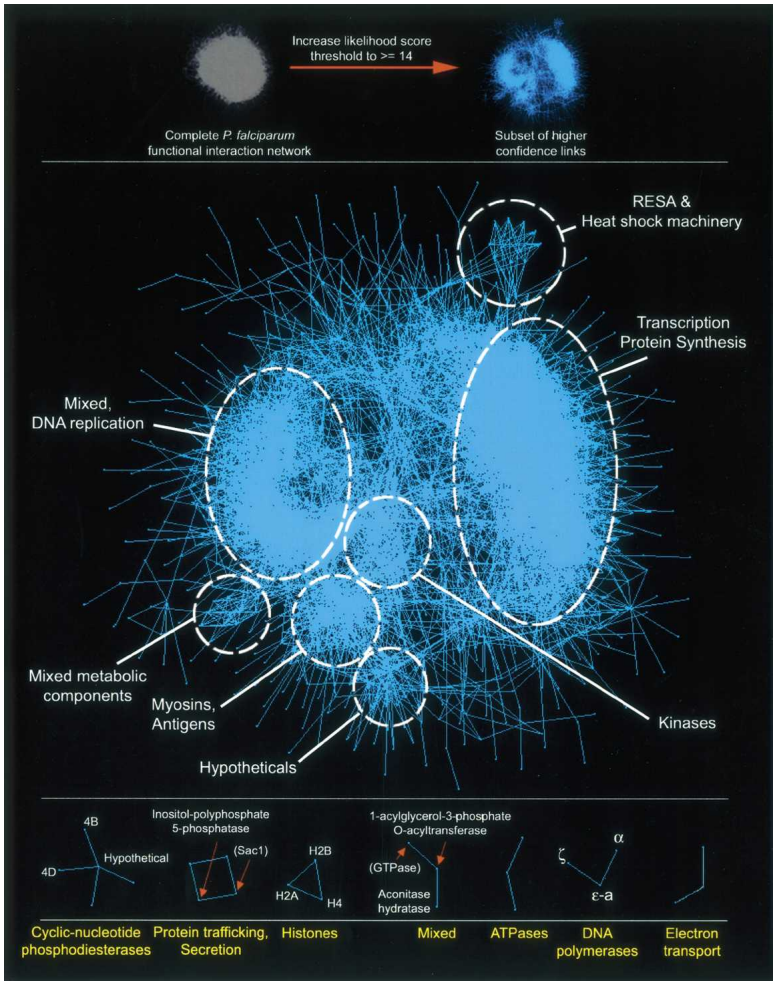


**Figure 4.** A subset of high confidence links as visualized by Large Graph Layout package (LGL, Adai et al. 2004). The subset of links at a score threshold of 14 or greater contains 10× the number of true positives than false positives when measured against our gold standards. Several functionally relevant subnetworks, such as that of proteins involved in transport, DNA polymerases, and histone proteins, are also illustrated. These subnetworks also reveal the various types of functionally relevant associations available; members that constitute parts of pathways, possess similar function, or even are subunits of protein complexes are visible. Edges in the graph are unweighted.

## Methods

### Assembling functional genomics data sets

The complete genome sequence of *P. falciparum* (Gardner et al. 2002) and ortholog information generated by OrthoMCL (Li et al. 2003) for *P. falciparum*, *P. yoelii* (Carlton et al. 2002), *T. gondii* (ToxoDB; http://toxodb.org), and *C. parvum* (Abrahamsen et al. 2004) were downloaded as available from PlasmoDB (versions 4.2 and 4.3; Kissinger et al. 2002). Amino acid sequences for 164 completely sequenced prokaryotic and eukaryotic genomes were downloaded from the National Center for Biotechnology Information Web site. Sequence comparisons between *P. falciparum* proteins and the downloaded genomes were used to construct phylogenetic profiles and determine the presence of Rosetta

**Table 2.** A *P. falciparum*-centric view of the Apicomplexan world

| | *P.f* only | *P.f* – *C.p* (1529) | *P.f* – *T.g* (2536) | *P.f* – *P.y* (4067) | *P.f* – All |
|---|---|---|---|---|---|
| Number of linkages | 87,798 | 60,748 | 138,252 | 287,931 | 47,364 |
| Number of proteins | 3307 | 1172 | 1978 | 3019 | 1001 |
| Number of unknown proteins | 2036 | 544 | 877 | 1783 | 420 |
| Number of shared pathways with ≥2 components | 68 | 43 | 67 | 68 | 38 |

The *P. falciparum* (*P.f*) functional interaction map was superimposed on the genomes of three other apicomplexan species: *C. parvum* (*C.p*), *T. gondii* (*T.g*), and *Plasmodium yoelii* (*P.y*). Numbers in parentheses in the header row indicate the number of orthologs detected by OrthoMCL (Li et al. 2003). These observations can be further refined as coverage and quality of the sequenced genomes improves.

stone fusion proteins. These data were filtered based on complexity of the profiles and the presence of Rosetta stone fusion proteins in other organisms (see Supplemental information for details). Genome-wide profile similarity was measured by calculating the mutual information for individual protein pairs (for details of this procedure, see Date and Marcotte 2003). Linkage confidence for each Rosetta stone pair was measured by calculating the probability of observing a given fusion by random chance, using a hypergeometric distribution, and a correction term based on the prevalence of homologs for proteins in a pair (Verjovsky Marcotte and Marcotte 2002).

Expression profiles for 3471 proteins, as published by Bozdech et al. (2003, HB3 strain) were obtained from PlasmoDB. These profiles were normalized using the "printTipLoess" method in R and smoothed using a least-squares fit to a sliding window of five points (see The R Project for Statistical Computing, http://www.r-project.org/; normalized profiles were kindly provided by G. Grant, Univ. of Pennsylvania) and Pearson correlation was calculated for all possible pairs over the entire data set. Smoothed and normalized expression profiles for the 3D7 and Dd2 strains were also obtained from PlasmoDB. These profiles represent average intensities of all probes that map to a particular gene for each time point.

### Data integration within a Bayesian framework

We created a set of gold standard reference sets using annotation from the KEGG database (Kanehisa et al. 2004) and GO process (The Gene Ontology Consortium 2004) annotation as supplied by the sequencing centers (see Gardner et al. 2002), available via PlasmoDB. The GO hierarchy was downloaded on December 16, 2004. Positive gold standard pairs ($G_P$) were created by pairing proteins within the same KEGG pathway, provided the proteins participated in three pathways or less, thus avoiding promiscuous members. Negative gold standards ($G_N$) were created by pairing all proteins, with KEGG information, and excluding pairs in the gold standard positive set. Exclusion of pairs that had one or more common GO terms in seven consecutively increasing levels of the GO hierarchy, allowed further filtration of this set of negative gold standards (see Supplemental information).

$G'_N = G_N \backslash S$, where S = {all protein pairs created based on one or more shared GO terms between the pair, in up to seven levels of the GO hierarchy}

For data integration within a naive Bayesian framework, we first ensured conditional independence of the in silico data sets, using scatter-plot analysis. Protein pairs from each of the functional genomics data sets were binned, based on the confidence measurements associated with each data set. For instance, expression profiles were binned for a 0.1 increase in correlation, whereas Rosetta stone links were binned for every $10^{-1}$ increase (see Supplemental Table 1). A likelihood ratio (LR) for each pair was calculated based on binned confidence measurements, as described by Jansen et al. (2003). Briefly, members of each bin were then tested for overlap with the gold standard sets, resulting in a likelihood ratio for each bin, which was then assigned to each member of the bin.

$$LR(Bin_{pairs}) = P(Bin_{pairs} \mid G_P) \;/\; P(Bin_{pairs} \mid G'_N)$$

A combined likelihood score for each protein pair was a product of the likelihood scores from each of the three data sets (*Phylo* = phylogenetic profile linkages, *Rosetta* = Rosetta stone linkages, *Expression* = gene expression profile linkages), with no penalties for missing evidence from any set. For proteins (*A*, *B*), the combined likelihood score is

$$LR(A,B) = LR(A,B)_{Phylo} \times LR(A,B)_{Rosetta} \times LR(A,B)_{Expression}$$

Likelihood score thresholds for each corresponding set of prior beliefs were computed based on Bayesian inference, as $O_{posterior} = O_{prior} \times LR$.

### Result testing and comparison

Initial quality of the generated predictions was measured by comparing links sampled from all possible pairwise linkages between *P. falciparum* proteins and two sets of randomly generated linkages. For the first round, we measured overlap of 100 random links from a set of all versus all pairs with the predicted functional linkages. One thousand repetitions of this test revealed an average overlap of 2.76 pairs. Next, we tested overlap of pairs generated using 100 random candidates (4590 pairs representing all versus all pairs within the 100 proteins) from proteins represented in the map and a shuffled set created by randomly pairing all included proteins, which generated the same number of pairs as in the map. These two tests reveal a false positive prediction rate of 5.80% and 12.85%, respectively. The average false positive prediction rate in our predictions, based on these three initial tests was ~7%.

Data from three large-scale studies—whole genome microarray-based expression profiles as measured by Le Roch et al. (2003) and two mass spectrometric measurements of the proteome (Florens et al. 2002; Lasonder et al. 2002)—were parsed to create sets of proteins that are present in one of three life-cycle stages: sporozoite stage, gamete/gametocytic stage, or other stages, which were collectively referred to as erythrocyte-associated stages for convenience. Proteins that shared the same or different life-cycle stages were paired to create positive and negative reference sets, respectively.

## Acknowledgments

We gratefully acknowledge J. Kissinger for suggestions and useful comments on the scientific content and organization of this manuscript and updates on missing genes in *C. parvum*. We would also like to thank D. Roos, J. Schug, H. He, E. Manduchi, and A. Ramani for their suggestions and comments. Mod-

ified expression profile data was kindly provided by G. Grant. C.J.S. is funded by the National Institutes for Health (grant 5R01AI058515).

## References

Abrahamsen, M.S., Templeton, T.J., Enomoto, S., Abrahante, J.E., Zhu, G., Lancto, C.A., Deng, M., Liu, C., Widmer, G., Tzipori, S., et al. 2004. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* **304:** 441–445.

Adai, A.T., Date, S.V., Wieland, S., and Marcotte, E.M. 2004. LGL: Creating a map of protein function with an algorithm for visualizing very large biological networks. *J. Mol. Biol.* **340:** 179–190.

Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J., and DeRisi, J.L. 2003. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.* **1:** 85–100.

Carlton, J.M., Angiuoli, S.V., Suh, B.B., Kooij, T.W., Pertea, M., Silva, J.C., Ermolaeva, M.D., Allen, J.E., Selengut, J.D., Koo, H.L., et al. 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii* yoelii. *Nature* **419:** 512–519.

Chen, F., Mackey, A., Stoeckert, C., and Roos, D. 2006. OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **34**: D363–D368.

Date, S.V. and Marcotte, E.M. 2003. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.* **21:** 1055–1062.

Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D., Moch, J.K., Muster, N., Sacci, J.B., Tabb, D.L., et al. 2002. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419:** 520–526.

Galat, A. and Metcalfe, S.M. 1995. Peptidylproline cis/trans isomerases. *Prog. Biophys. Mol. Biol.* **63:** 67–118.

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

Gavigan, C.S., Kiely, S.P., Hirtzlin, J., and Bell, A. 2003. Cyclosporin-binding proteins of *Plasmodium falciparum*. *Int. J. Parasitol.* **33:** 987–996.

The Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32:** D258–D261.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. 2003. A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* **302:** 449–453.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32:** D277–D280.

Kissinger, J.C., Brunk, B.P., Crabtree, J., Fraunholz, M.J., Gajria, B., Milgram, A.J., Pearson, D.S., Schug, J., Bahl, A., Diskin, S.J., et al. 2002. The Plasmodium Genome Database. *Nature* **419:** 490–492.

Lasonder, E., Ishihama, Y., Andersen, J.S., Vermunt, A.M., Pain, A., Sauerwein, R.W., Eling, W.M., Hall, N., Waters, A.P., Stunnenberg, H.G., et al. 2002. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419:** 537–542.

Lee, I., Date, S.V., Adai, A.T., and Marcotte, E.M. 2004. A probabilistic functional network of yeast genes. *Science* **306:** 1555–1558.

Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holder, A.A., Batalov, S., Carucci, D.J., et al. 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301:** 1503–1508.

Li, L., Stoeckert Jr., C.J., and Roos, D.S. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13:** 2178–2189.

Marcotte, E.M., Pellegrini, M., Ng, H.-L., Rice, D.W., Yeates, T.O., and Eisenberg, D. 1999. Detecting protein function and protein–protein interactions from genome sequences. *Science* **285:** 751–753.

Marti, M., Good, R.T., Rug, M., Knuepfer, E., and Cowman, A.F. 2004. Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306:** 1930–1933.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* **96:** 4285–4288.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32:** D449–D451.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **11:** 2498–2504.

Silva, M.D., Cooke, B.M., Guillotte, M., Buckingham, D.W., Sauzet, J.P., Le Scanf, C., Contamin, H., David, P., Mercereau-Puijalon, O., and Bonnefoy, S. 2005. A role for the *Plasmodium falciparum* RESA protein in resistance against heat shock demonstrated using gene disruption. *Mol. Microbiol.* **56:** 990–1003.

Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., and Botstein, D. 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *S. cerevisiae*). *Proc. Natl. Acad. Sci.* **100:** 8348–8353.

Verjovsky Marcotte, C.J. and Marcotte, E.M. 2002. Predicting functional linkages from gene fusions with confidence. *App. Bioinform.* **1:** 1–8.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., and Bork, P. 2002. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417:** 399–403.

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

*Received August 17, 2005; accepted in revised form December 22, 2005.*