Predicting Malaria Interactome Classifications from Time-Course Transcriptomic Data along the Intraerythrocytic Developmental Cycle

Antonina Mitrofanova [1]*, Samantha Kleinberg [1], Jane Carlton [2],
Simon Kasif [3], Bud Mishra [1],

[1] Computer Science Department, Courant Institute of Mathematical Sciences,
New York University, 715 Broadway, New York, NY 10003

[2] Department of Medical Parasitology, New York University,
550 Fifth Avenue, New York, NY, 10016

[3] Department of Biomedical Engineering, Boston University,
44 Cummington Street, Boston, MA, 02215

August 4, 2009

Summary

Objective: Even though a vaccine for malaria infections has been under intense study for many years, it has resisted several different lines of attack attempted by biologists. More than half of Plasmodium proteins still remain uncharacterized and therefore cannot be used in clinical trials. The task is further complicated by the metamorphic life cycle of the parasite, which allows for rapid evolutionary changes and diversity among related strains, thus making precise targeting of the appropriate proteins for vaccination a technical challenge. We propose an automated method for predicting functions for the malaria parasite, which capitalizes on the importance of the intraerythrocytic developmental cycle data and expression changes during its five phases, as determined computationally by our segmentation algorithm.

Materials and methods: Our method combines temporal gene expression profiles with protein-protein interaction data, sequence similarity scores, and metabolic pathway information to produce a set of predicted protein functions that can be used as targets for vaccine development. We use a Bayesian approach, which assigns a probability of having (or not having) a particular function to each protein, given the various sources of evidence. In our method, each data source is represented by either a functional linkage graph or a categorical feature vector.

*to whom correspondence should be addressed, e-mail: $antonina@cs.nyu.edu$, phone number: 1-212-998-3374, fax number: 1-212-998-3484

Results and conclusions: The methods are tested on *Plasmodium falciparum*, the species responsible for the deadliest malaria infections. The algorithm was able to assign meaningful functions to 628 out of 1439 previously unannotated proteins, which are first-choice candidates for experimental vaccine research. We conclude that analyzing time-course gene expression profiles in separate phases leads to much higher prediction accuracy when compared with Pearson correlation coefficients computed across the time course as a whole. Additionally, we demonstrate that temporal expression profiles alone are able to improve the predictive power of the integrated data.

Keywords: Plasmodium falciparum, intraerythrocytic developmental cycle, time-course gene expression data, Bayesian probabilistic approach, protein function prediction, RBC membrane proteins, HT motif, Pexel.

1 Introduction

1.1 Background

World-wide, each year, malaria infects approximately 515 million people and kills between one and three million of them. A better understanding of *protein functions* in malaria parasites can have a tremendous effect on approaches aimed at preventing malaria epidemics. This anticipated impact is suggested by the fact that targets for drug and vaccine design are almost always based on proteins, particularly those involving enzymatic functions. Unfortunately, since many *Plasmodium falciparum* proteins remain uncharacterized, they are mostly ignored by pharmaceutical laboratories and disregarded as potential protein targets in drug and vaccine development. In order to reverse this trend, it is necessary to devise more effective automated bioinformatic tools for protein classification.

Toward this goal, this paper addresses the issue of predicting protein functions using many sources of data, with an emphasis on the use of time series gene expression data. Unlike most methods, we allow for changes in regulatory patterns, and relationships, over time. The methods are tested on a species of malaria parasite, *P. falciparum*, that accounts for about 15% of infections and 90% of deaths.

In the past, functional annotation of proteins has been addressed by various computational, statistical, and experimental methods. In many cases, it is convenient to provide a graphical representation of protein networks such that each node represents a protein and edges between nodes represent different aspects of their functional association. The choice of functional association is used to determine the predictive power of such a network. One promising computational approach utilizes the family of probabilistic graphical models, such as belief networks, to infer functions over sets of partially annotated proteins [1–4]. For instance, Bayesian network methods for data integration have been extensively studied [5–8] for predicting protein-protein interactions and protein function similarity for pairs of genes. Additionally, the approach of incorporating the hierarchical structure of the Gene Ontology (GO) into probabilistic graphical models [9, 10] has also yielded promising results for predicting protein functions for gene subontologies of interest.

The most established methods for protein function prediction are based on sequence similarity using BLAST [11] analysis, and rely on the fact that similar proteins are likely to share common functions. Such similarity-based methods include sequence homology [9, 10, 12–15], and similarity in short signaling motifs, amino acid composition and expression data [16–21]. At the same time, protein-protein interaction (PPI) data are widely used to infer protein functions. For example, methods described in several recent papers [1–4] used PPIs to define a Markov random field over the entire set of proteins. In general these methods suggests that interacting neighbors in PPI networks might also share a function [1, 22–24]. Clustering of genome-wide expression patterns has also been used to predict protein function, as described in [5, 25–27].

1.2 Protein function prediction in parasites

Saccharomyces cerevisiae (Baker's Yeast) is chosen for many case studies, since it has been extensively studied from multi-omic view-points, and its protein data are also the most complete. The problem of protein function prediction is, however, more difficult in parasites, where genetic and biochemical investigations are much more challenging. For example, it is problematic to isolate a malaria parasite at various stages of its development (e.g., the life-cycle of *P. falciparum* is very rapid, ookinetes are difficult to isolate in large numbers, the liver stage of a parasite's development is hard to study because of technical difficulties). Such obstacles manifest themselves in a paucity of information on the protein properties, interactions, localization and motifs of *Plasmodium* species.

When relying on just one source of protein information, it is difficult to devise a reliable probabilistic framework with the ability to automatically predict classifications for proteins of interest. Indeed, combining various types of information was demonstrated to improve the overall predictive power of automated protein/gene annotation systems for *S. cerevisiae*, as shown in [5, 9, 10]. Integrating multiple sources of information is particularly important as each type of data captures only one aspect of cellular activity. For example, PPI data suggest a physical interaction between proteins; sequence similarity captures evolutionary relationships at the level of orthologs; gene expression suggests participation in related biological processes that take place at a certain cell cycle stage; and finally, GO defines term-specific dependencies.

As a result, it motivates one to explore, as in the case of *P. falciparum*, how to combine different sources of information most effectively to infer protein functions. We explore and evaluate a Bayesian probabilistic approach for predicting protein functions in *P. falciparum* by integrating multiple sources of information, namely, protein-protein interactions, sequence similarity, temporal gene expression profiling, metabolic pathway, and GO classifications.

The primary goal of our study is to demonstrate that considering the intraerythrocytic developmental cycle (IDC) phases individually is crucial for protein function prediction in *P. falciparum*. While other data sources (such as sequence homology and protein-protein interactions) describe the static state of *P. falciparum*, time series gene expression data during the IDC reflects the dynamics of the parasite's system, describing rapidly evolving regulatory patterns and expression profiles. In particular, during *P. falciparum*'s IDC, there are distinct periods of consistent gene regulation, punctuated by instances of reorganization in the regulation pattern. In such a setting, it becomes important to consider each time window (delineating a particular stage) separately. We show that finding these critical timepoints, clustering time-course gene expression data from each stage of the cycle *separately* and then connecting clusters across windows (so that proteins "travel" from one window to the other) produces better results as compared with Pearson coefficient calculations applied to the time-course data as a whole. We assume that if two proteins share expression patterns (i.e., belong to the same cluster) during a period of time, such as the first window or phase, they are likely to share a function. If these proteins also fall into the same cluster in the second window, we would increase our belief in them being similar. Finally, if they belong to the same clusters in all five windows, we would be highly confident that they share related functions.

Additionally, but not less importantly, we illustrate that inclusion of the IDC time-course data improves the predictive power of the Bayesian probabilistic approach even in the integrated setting (when combined with protein-protein interaction, sequence homology, and metabolic pathways data).

Hampered by data-related limitations, we did not expect to make as many accurate predictions as one could for a well-studied organism such as *S. cerevisiae*. However, we were encouraged by being able to propose vaccine-related functions for several *P. falciparum* proteins as these might play a significant role in the next stages of vaccine and drug development, leading to effective control of the disease.

The next part of this process involves trying to understand the underlying causal structure that is governing *P. falciparum*'s gene regulation. That is, now that we have a set of possible functional annotations, and time course data covering the IDC, we aim to narrow the set of proteins suitable for vaccine exploration by finding those that can be used to affect others. Note that it is likely not as simple as one protein promoting or inhibiting the production of another - there may be arbitrarily complex relationships involving the regulation of multiple genes in concert. We have previously developed algorithms for causal inference, where the relationships are described in a probabilistic temporal logic, allowing arbitrarily complex causes and effects and explicit description of the time between the cause and the effect. Preliminary results of the *P. falciparum* IDC have appeared elsewhere [28]. One of the limitations of this data is the relatively coarse timescale (as compared to other data sets used for causal inference). Rather than exhaustively examining all proteins included in the data, we now plan to focus on a smaller set of relationships to be tested, using our new annotations and processes known to be useful as drug targets.

2 Methods

2.1 Data

For our analysis, we focused on 2688 *P. falciparum* proteins from the time-course data [29], among which only 1249 proteins possess known biological process annotations.

2.1.1 Protein-protein interaction data

We obtained Y2H (yeast-two-hybrid) data for *P. falciparum* from [30]. This dataset, however, annotates a limited number of protein-protein interactions, because of the confounding effects of the rapid life-cycle of these parasites. The 1130 interactions cover 1312 proteins.

2.1.2 Sequence homology

We started by gathering sequence information for proteins from [30]. Each sequence was queried against the entire *P. falciparum* sequence database [30] using BLAST. We recorded BLAST pairwise p-scores as p_{ij}'s (where i and j index the proteins) and defined a measure of sequence similarity for each pair as $s_{ij} = 1 - p_{ij}$. For our purpose, we defined proteins i and j to be similar (sequence-wise), if their pairwise p-value $p_{ij} < 10^{-4}$. There are 1799 proteins meeting this criteria.

2.1.3 Metabolic pathway data

We used metabolic pathway data from [31]. For example, protein PFA0145c is a part of 'Asparagine and Aspartate metabolism' and 'Protein biosynthesis' pathways. The data consists of 119 metabolic pathway categories for *P. falciparum*. The 3526 data pairs cover 1998 genes.

2.1.4 Temporal gene expression data

Time-course gene expression data covering the 48 hours of the intraerythrocytic developmental cycle of *P. falciparum* was obtained from a study by Bozdech et al. [29]. While the IDC comprises three main stages (ring, trophozoite, and schizont, separated by two critical transition instants), the work in [32] identified four critical transition instants with major changes in gene regulation, corresponding to the following five developmental periods: End Merozoite/Early Ring stage, Late Ring stage/ Early Trophozoite stage, Trophozoite, Late Trophozoite/ Schizont, and Late Schizont/Merozoite. Each period defines a window of time ranging from 7 to 16 hours. We consider each window separately and process it with k-means clustering.

2.1.5 Gene Ontology data

We used GO terms as the basis of our annotation. In particular, we used the 763 biological process associated GO terms available for *P. falciparum*. For each term we expanded the GO hierarchy "up" (including is-a and part-of relationships) so that if a protein is positively annotated by a GO term, then it is also positively annotated to all of its parents/ancestors. There are 16113 GO biological process associated pairs, which cover 1249 *P. falciparum* proteins. Following Nariai et al. [5], we excluded labels that appear fewer than five times among these genes, since these terms did not constitute a sample large enough to make sufficiently predictive contributions. Following suggestions in Nariai et al. [5], we define a negative protein-term association as follows: if the association is not in the positive set (defined above), and a gene is annotated with at least one biological process, and the negative annotation is neither an ancestor nor a descendant of the known function for this protein then it is treated as a negative association.

2.2 Data representation

In order to use the available information to its full potential, it is necessary to design a proper data representation that optimally reflects the properties and structure of the data itself. We represent the data using two types of structures: *functional linkage graphs* and *categorical feature vectors*.

A *functional linkage graph* is a network in which each node corresponds to a protein and each edge corresponds to the measure of functional association. Such a network takes into account the number and the nature of interacting partners for each protein. We use this representation for PPI and sequence similarity, since, for these data, interacting partners are more likely to share a function. We encoded PPI and sequence homology data using separate functional linkage graphs. In the case of PPI, the edges represent known protein-protein interactions. In the case of sequence similarity (homology) an edge is added when the pairwise p-score is less than 10^{-4}.

We adopted some ideas of the representation and analysis of functional linkage graphs from Nariai et al. [5]. For each functional linkage graph l and for each GO label t, we define $p_1^{(l)}$ and $p_0^{(l)}$, where $p_1^{(l)}$ is the probability that a protein has label t, given that the interacting partner has label t and $p_0^{(l)}$ is the probability that a protein has label t given that the interacting partner does not have label t. For the *P. falciparum* network, we used the χ^2 test to show that these probabilities are statistically different and used a Bonferroni-corrected p-value of $0.001/T$, where T is the number of terms tested from each data set.

Another method of data representation is the *categorical feature vector*, which holds a list of categories where we assign 1 to a protein that belongs to a given category and 0 otherwise. We used categorical feature vectors for the metabolic pathway data. We define m_r as a random variable associated with a protein so that $m_r = 1$, if a protein participates in metabolic pathway r, and $m_r = 0$, otherwise. A feature vector $\mathbf{m} = (m_1, m_2, \ldots, m_r)^T$ is defined for each protein, where $r = 119$ is the total number of metabolic pathway categories.

Finally, we use categorical feature vectors to represent the gene expression profiles. Gene expression profiles are usually encoded as functional linkage graphs using the Pearson correlation coefficient calculated for all combinations of genes. However, we found that the Pearson coefficient might not reflect the temporal relationships, which are crucial to the *P. falciparum* IDC. Instead, we consider expression data for each phase of the IDC separately. We used the five time points found by [32] and applied k-means clustering to the expression patterns of each time period, as described below. We considered proteins from the same cluster to share the same categorical feature and thus possibly have related functional annotations. Consequently, if proteins fall into the same clusters for all or most of the time periods, they will have similar categorical feature vectors and are more likely to share protein classification.

More formally, we define a random variable d_r^j associated with each protein such that $d_r^j = 1$ if a protein is in cluster r in the time period j, and $d_r^j = 0$, otherwise. A feature vector is then

$$\mathbf{d} = (d_1^1, d_2^1, \ldots, d_q^1, d_1^2, d_2^2, \ldots, d_q^2, \ldots, d_1^w, d_2^w, \ldots, d_q^w)^T,$$

where $q = k$ is the number of clusters produced by k-means clustering and $w = 5$ is the number time windows.

2.3 Posterior probability computation

For each protein i and each function t, we computed the posterior probability of the protein having the specified function. We define a variable $L_{i,t}$ which is equal to 1 if i is labeled with t. Our ultimate goal is to calculate the probability of $L_{i,t} = 1$ for all i and t given all the available data sources and network structures. To calculate this probability, we follow the general principles described in Nariai et al. [5] and summarize these principles below.

The graphical data representation emphasizes the importance of the neighbors for each protein. We define $N_i^{(l)}$ as the number of neighbors of protein i in the functional linkage graph l (unannotated neighbors are excluded). Additionally, for the corresponding t, $k_i^{(l)}$ is defined as the number of neighbors of protein i annotated with term t in the functional linkage graph l. In our case, $l = 1$ corresponds to the PPI and $l = 2$ to the sequence similarity network, .

At the same time, $\mathbf{c}_i^{(j)}$ is the feature vector that protein i has for a functional category j. In our case, $\mathbf{c}_i^{(1)}$ is the temporal data gene expression feature vector **d** and $\mathbf{c}_i^{(2)}$ is a metabolic pathway feature vector **m** of a protein i.

We calculate the posterior probability of $L_{i,t} = 1$ given functional linkage graphs and category feature vectors of proteins as follows:

$$\begin{aligned}
&P(L_{i,t} = 1 | N_i^{(1)}, k_i^{(1)}, N_i^{(2)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)}) && (1)\\
&= \frac{P(L, N_i^{(1)}, k_i^{(1)}, N_i^{(2)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)})}{P(N_i^{(1)}, k_i^{(1)}, N_i^{(2)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)})} && (2)\\
&= \frac{P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)} | L, N_i^{(1)}, N_i^{(2)}) P(L, N_i^{(1)}, N_i^{(2)})}{P(N_i^{(1)}, k_i^{(1)}, N_i^{(2)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)})} && (3)\\
&= \frac{P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)} | L, N_i^{(1)}, N_i^{(2)}) P(L | N_i^{(1)}, N_i^{(2)})}{P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)} | N_i^{(1)}, N_i^{(2)})} && (4)
\end{aligned}$$

Assuming that k's and **c**'s are independent, and that L is independent of the total number of graph neighbors $N_i^{(l)}$, then the numerator becomes:

$$\begin{aligned}
&P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)} | L, N_i^{(1)}, N_i^{(2)}) P(L | N_i^{(1)}, N_i^{(2)}) && (5)\\
&= \prod_{l=1}^{2} P(k_i^{(l)} | L, N_i^{(1)}, N_i^{(2)}). \prod_{j=1}^{2} P(\mathbf{c}_i^{(j)}, | L, N_i^{(1)}, N_i^{(2)}) \times P(L), && (6)
\end{aligned}$$

and similarly, the denominator becomes

$$
\begin{aligned}
& P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)} | N_i^{(1)}, N_i^{(2)}) && (7)\\
& = P(L) P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)} | L, N_i^{(1)}, N_i^{(2)}) && (8)\\
& + P(\neg L) P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)} | \neg L, N_i^{(1)}, N_i^{(2)}). && (9)
\end{aligned}
$$

Assuming further that $k_i^{(l)}$ only depends on $N_i^{(l)}$ and that $\mathbf{c}_i^{(j)}$ does not depend on linkage graphs,

$$
\begin{aligned}
& \prod_{l=1}^{2} P(k_i^{(l)} | L, N_i^{(1)}, N_i^{(2)}). \prod_{j=1}^{2} P(\mathbf{c}_i^{(j)}, | L, N_i^{(1)}, N_i^{(2)}) \times P(L) && (10)\\
& = \prod_{l=1}^{2} P(k_i^{(l)} | L, N_i^{(l)}) \times \prod_{j=1}^{2} P(\mathbf{c}_i^{(j)} | L) \times P(L), && (11)
\end{aligned}
$$

and

$$
\begin{aligned}
& P(L) P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)} | L, N_i^{(1)}, N_i^{(2)}) && (12)\\
& + P(\neg L) P(k_i^{(1)}, k_i^{(2)}, \mathbf{c}_i^{(1)}, \mathbf{c}_i^{(2)} | \neg L, N_i^{(1)}, N_i^{(2)}) && (13)\\
& = P(L) \prod_{l=1}^{2} P(k_i^{(l)} | L, N_i^{(l)}) \prod_{j=1}^{2} P(\mathbf{c}_i^{(j)} | L) && (14)\\
& + P(\neg L) \prod_{l=1}^{2} P(k_i^{(l)} | \neg L, N_i^{(l)}) \prod_{j=1}^{2} P(\mathbf{c}_i^{(j)} | \neg L). && (15)
\end{aligned}
$$

Similarly to the other formulations in the literature [1, 5], $P(k_i^{(l)} | L, N_i^{(l)})$ and $P(k_i^{(l)} | \neg L, N_i^{(l)})$ are calculated assuming the binomial distribution. $P(L)$ is the prior probability that gene i is annotated with term t and is calculated as a frequency of term t among genes.

3 Experiments and results

For the 5-fold cross-validation study, we created each test set by eliminating all annotations from a random 20% of annotated proteins (250 randomly chosen proteins from the annotated set of 1249). We performed 5 validation runs and report the average of these for the summary statistics. We use the statistical measures $sensitivity$ and $specificity$, as defined in [33]. We also use the $F1$ measure which represents a weighted harmonic mean of precision and recall and is defined as

$$
F1 = \frac{2 \times (precision \times recall)}{precision + recall}
$$

Note that $F1$ allows analysis of the performance weighing precision and recall evenly.

3.1 Gene expression data of a parasite life-cycle

First, we show and emphasize the importance of gene expression data representation and analysis, especially when applied to parasites. Many parasites, such as malaria parasites, trypanosomes, endoparasites with larval stages (tapeworms, thorny-headed worms, flukes, parasitic roundworms), undergo many changes during their various life-cycle stages as they travel from one host to the other, or from one organ or system to another. Each stage requires utilization of different life functions and possible metamorphosis, which involves up-regulation of necessary genes and/or down-regulation of those not crucial for a specific life-cycle period.

In this study we use the five time windows of the intraerythrocytic developmental cycle (IDC) of *P. falciparum* identified by Kleinberg et al. [32]. This expression data is particularly interesting since the IDC, or blood stage, is the phase responsible for malaria symptoms in humans. This study [32] performs the time series segmentation and clustering of the data concurrently. Their method is formulated in terms of rate distortion theory—it searches for a compressed description of the data (i.e. the fewest clusters of expression profiles, obtained after an optimal temporal segmentation), while minimizing the distortion introduced by this compression. More formally, this process is characterized by a variational formulation:

$$\mathcal{F}_{min} = I(Z;X) + \beta\langle d(x,z)\rangle, \tag{16}$$

where mutual information and average distortion are defined as:

$$I(Z;X) = \sum_{x,z} p(z|x)p(x)log\frac{p(z|x)}{p(z)} \tag{17}$$

$$\langle d(x,z)\rangle = \sum_{x,z} p(x)p(z|x)d(x,z), \tag{18}$$

and

$$d(x,z) = \sum_{x_1} p(x_1|z)d(x_1,x). \tag{19}$$

Then, the set of candidate windows (i.e., enumeration of all possible windowings within constraints on the min and max allowed window sizes) is created, and the data is clustered within each window according to Eq. (16). Each window is then scored, based on its length and Eq. (16). To find the optimal windowing of the data, they formulate the problem as one of graph search and use a shortest path algorithm to find a combination of windows that jointly provide the lowest cost. For the *P. falciparum* data the study in Kleinberg et al. [32] found the critical time points at 7, 16, 28 and 43 hours, leading to 5 windows, sized non-uniformly. These windows correspond to the three IDC stages and the transitions between them: End Merozoite/Early Ring stage, Late Ring stage/ Early Trophozoite stage, Trophozoite, Late Trophozoite/ Schizont, and Late Schizont/Merozoite.

The clustering by Kleinberg et al. [32] identified 4-5 clusters per window, corresponding to the three phases of the cycle with an additional one or two clusters per window containing terms regulating the beginning or end of a phase. In order to predict detailed functional annotations, we decided to cluster the data more finely. We use these previously identified windows and clustered the expression profiles within each separately, using the k-means clustering algorithm. We then define d_r^j as a random variable indicating if a protein belongs in the cluster r within window j. The sequence of random variables for each window then constitutes a categorical feature vector **d** of a protein.

We experimented with various values for k and compared results with the linkage graph defined by a Pearson coefficient calculation; we performed this step for all pairs of genes for the entire data set.

In our experiments, due to a high number of negative annotations for the *P. falciparum* dataset, specificity reaches 0.9 immediately after the threshold for posterior probability goes above 0.05. In this case a ROC curve, as shown in Figure 1, does not reflect a precise sensitivity-specificity relationship as expected in other cases, obtained with a relatively large amount of data. As a result, it is necessary to use a more sensitive statistical measure that would account for too high or too low statistical values, e.g., a metric computed by taking their harmonic mean. In particular, we aim to maximize the $F1$ statistic, which reflects a relationship of recall to precision, as noted in Figure 2. Note that $F1$ will be maximized only if both measures are maximized.

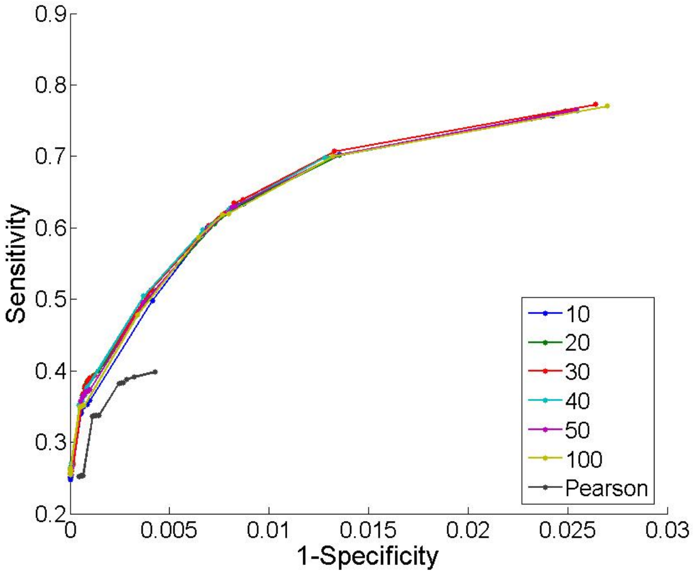

Figure 1: The ROC curve of recall experiment by 5-fold cross validation for gene expression data. Numbered legends correspond to k-means clustered datasets.

As shown in Figures 2, the variation in the number of clusters, k, does not distort the predictive value of the method as for all values of k in this range, the method yields nearly identical ROC and F1 curves. Figure 2 also shows the superiority of time-dependent k-means clustering over the Pearson correlation coefficient (in the majority

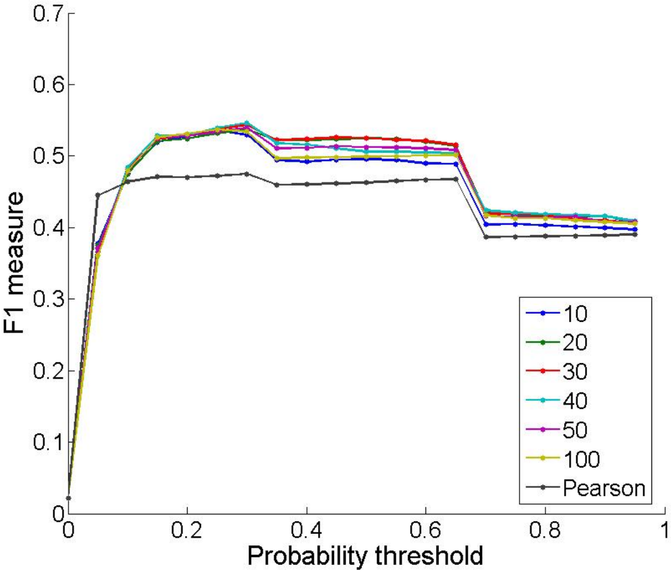

Figure 2: The F1 statistics of recall experiment by 5-fold cross validation for gene expression data (posterior probability thresholds range from 0.05 to 0.95, in 0.05 increments). Numbered legends correspond to k-means clustered datasets.

of cases the Pearson curve is completely below the curves for the clustered data). The linkage graph defined by the Pearson correlation coefficient was built using 286620 edges (a protein pair is considered co-expressed if its Pearson coefficient is larger than 0.85 [5]) and covered 2646 proteins.

Since for all values of k both figures showed nearly identical ROC and F1 curves, we fixed it at an arbitrary value, $k = 30$, for the following analysis.

3.2 Analysis of prediction accuracy

We compare runs on individual data sources with runs which integrate PPI, sequence similarity, metabolic pathway information, and temporal gene expression data. Our first step is to analyze how well our method predicts known protein-term associations, using 5-fold cross validation. We predict that a gene i is annotated with term t if the probability exceeds a specified threshold.

Figures 3and 4 summarize the positive impact of data integration (PPI, sequence similarity, metabolic pathway, window-based gene expression clustering) on protein function prediction via ROC and $F1$ measures, respectively. However, since ROC curves are very much influenced by the large number of negative annotations in *P. falciparum* data (similarly to Figure 1), specificity reaches 0.9 immediately after the threshold for posterior probability goes above 0.05), this measure is not very sensitive with respect to specificity scores; thus, we prefer the F1 statistic, which uses the harmonic mean of precision and recall. In these figures, we also show the statistics for the data obtained by analysis of gene expression using Pearson correlation coefficients (showing a clear disadvantage), although it was not a part of the data integration.

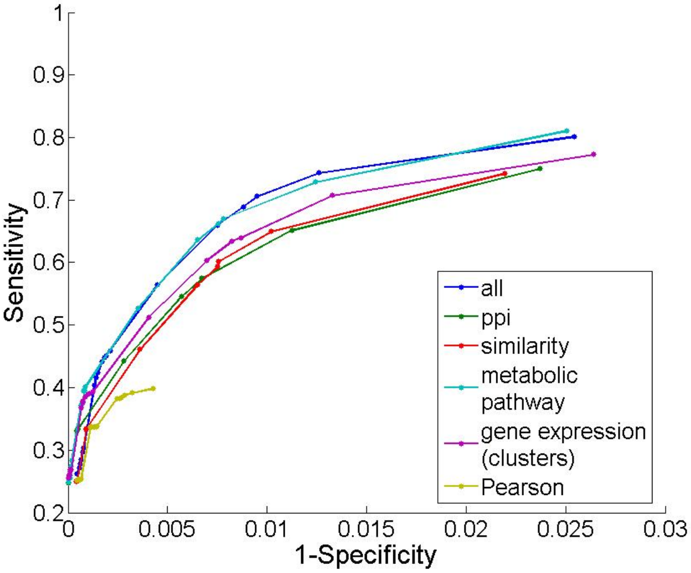

Figure 3: The ROC curves for individual data sources and integrated data.

Additionally, we investigated the impact of gene expression data on data integration. In Figures 5 and 6, we show both ROC and F1 curves, respectively, for fused data (PPI, similarity, and metabolic pathway) alone, then for fused data together with the windowed and clustered gene expression data, and fused data with Pearson coefficient defined data. Clustered temporal gene expression data shows a distinctive positive impact on the overall predictive power of the method; however, Pearson coefficient data has a negative effect on ROC and F1 statistics. Most likely this anomaly is due to a large number of falsely defined associations between co-expressed genes.

Figure 7 shows the impact of data integration on the number of TP at two precision levels: 50% and 70%. These two levels of precision are reasonably accurate of the range of possible improvements in our study, and the TP number is calculated when the precision level first hits the specified margin. In Table 1, we summarize the improvements of data integration over individual sources and conclude that data integration significantly outperforms individual data sources at 70% precision, which corresponds to 0.35 threshold of posterior probability for function prediction. This probability threshold now can be applied in the second step of our study: attempting to predict functions for the unannotated proteins of *P. falciparum*.

In the second part of our study, we trained our method on all annotated proteins

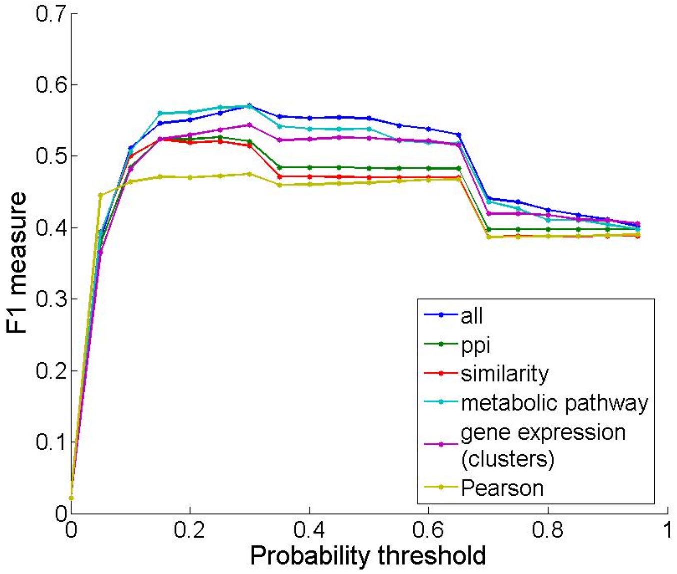

Figure 4: The F1 statistics for individual data sources and integrated data (posterior probability thresholds ranges from 0.05 to 0.95, in 0.05 increments).

Data source	50% precision	70% precision
PPI	14%	20%
Sequence similarity	17%	23%
Metabolic pathway	5%	13%
Gene expression (clustering)	11%	10%

Table 1: % of improvements of data integration on #TP over individual data sources

and tried to assign functions to proteins without annotations. By integrating PPI data, sequence similarity, metabolic pathway, and clustered temporal window-based gene expression data we were able to assign probable GO terms to 628 out of 1439 unannotated proteins of *P. falciparum*. We ignored general terms, such as those high up in the GO hierarchy, that appeared more than 300 times. We report more than 2500 gene-GO assignment pairs, which can be viewed at: `http://www.cims.nyu.edu/~antonina/real_output.txt`. The GO terms are reported together with their parents (ancestors) in the GO hierarchy. In Figure 8, we present cumulative statistics for the number of predicted functional assignments and probability thresholds they satisfy. As shown in Figure 8, by varying the original probability threshold, we can narrow down the possible set of predictions. For example, probability threshold at 0.8 (80%) would correspond to about 500 functional assignments of higher probability.

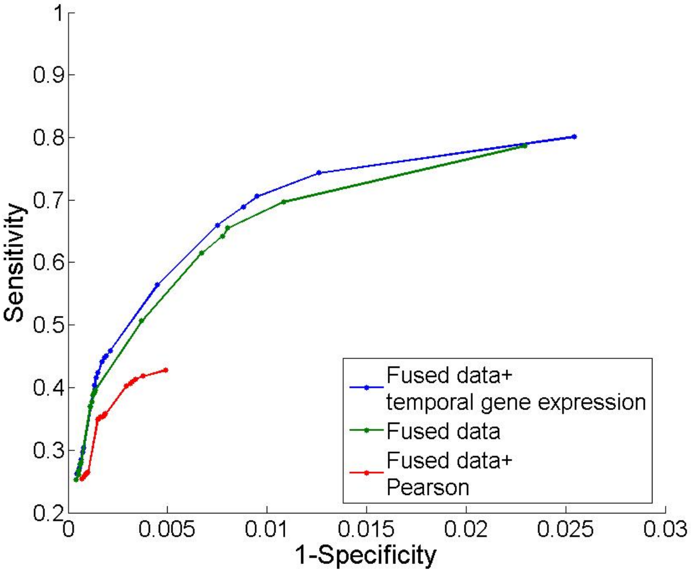

Figure 5: The ROC curves for various ways of integrating data:“fused” is defined as ppi+similarity+metabolic pathway.

4 Functional predictions for pharmaceutical targeting

The fundamental goal of our study is to assign functions to unannotated *Plasmodium falciparum* protein hoping to find possible vaccine and drug targets. For this purpose, we analyze all predicted functional assignments made by our computational technique for being related to erythrocytic adhesion and modification. In particular, we pay a special attention to *Plasmodium falciparum* surface proteins responsible for binding of the parasite to human erythrocytes, and to the *Plasmodium falciparum* red blood cell (RBC) membrane proteins responsible for parasite’s intraerythrocytic survival and for the adhesion of the RBC to capillary vessels. In our predicted dataset of 628 proteins, 20 are identified as RBC membrane proteins (contributing to 78 functional predictions) and one protein is identified as erythrocyte binding proteins (contributing to two functional predictions).

We further classify RBC membrane proteins as those having one of the address tags: either *Plasmodium export element* (Pexel) or *N-terminal host targeting* (HT) motif. Both of these motifs are responsible for a transport of *Plasmodium falciparum* proteins inside erythrocytic cytoplasm, as detailed below.

During malaria blood stage, *Plasmodium falciparum* actively penetrates human erythrocytes. In the process of invasion, the parasite initiates the formation of a unique membrane, the parasitophorous vacuole membrane, which surrounds the parasite inside the invaded erythrocyte. The parasitophorous vacuole isolates the parasite and

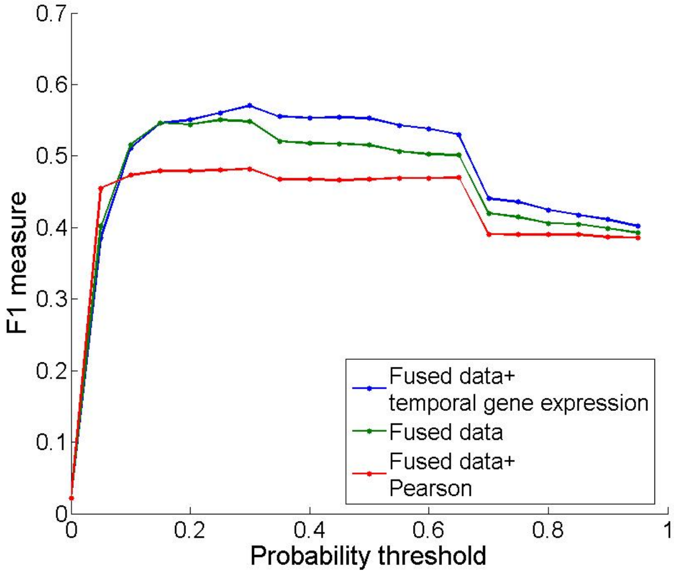

Figure 6: The F1 statistics for various ways of integrating data:“fused” is defined as ppi+similarity+metabolic pathway (posterior probability thresholds ranges from 0.05 to 0.95, in 0.05 increments).

protects it from the host’s defenses, such as lysozymal attack.

Plasmodium falciparum needs to develop its own strategy in order to survive and feed inside human erythrocytes since red blood cells lose nucleus, ability to synthesize new proteins, and a vesicular transport system during their formation. Residing inside a red blood cell, *Plasmodium falciparum* injects hundreds of its own proteins into erythrocytic cytoplasm [34] to build its living environment. The injected proteins then interact with proteins of the erythrocytic membrane skeleton and induce substantial changes in the morphology and function of the erythrocytic cell. Such changes include development of various membraneous (tubulovesicular structures and Maurer’s clefts) networks from the vacuole to the erythrocyte membrane, which are needed for parasite’s nutrient uptake, and protrusion of the erythrocyte membrane in a form of electron-dense elevations called adhesive knobs [35].

To reach the erythrocytic cytoplasm and membrane, *Plasmodium falciparum* exported proteins have to traverse a series of physical barriers: parasite membrane, parasitophorous vacuole membrane, and sometimes erythrocytic membrane [35, 36]. First, proteins are exported form the parasite into the vacuolar space following the typical secretion pathway existing in all eucaryotic cells. However, a special mechanism is needed to cross the parasitophorous vacuole membrane and reach the erythrocytic cytoplasm. For the majority of *Plasmodium falciparum* proteins, an N-terminus host targeting (HT) motif [36, 37] is required to cross the vacuole membrane.

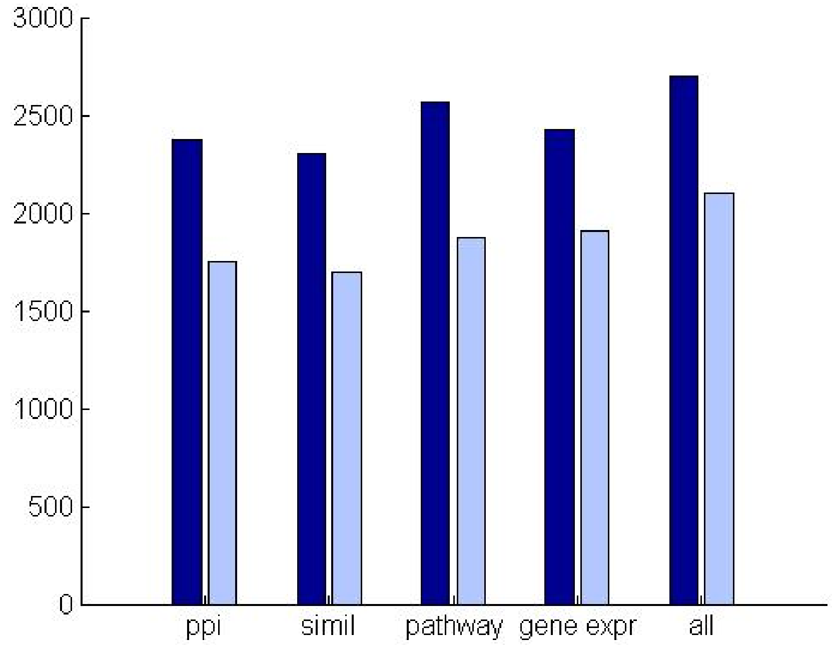

Figure 7: Number of True Positive predictions at 50% precision (dark blue) and at 70% precision (light blue)

On the other hand, Pexel is a Plasmodium export element (related but distinct from HT), responsible for the transport of only soluble *Plasmodium falciparum* proteins into erythrocyte cytoplasm through the parasitophoros vacuole membrane [36].

Exported proteins then interact with erythrocytic membrane causing its deformation and knobbing. Knobs mediate cytoadherence of infected erythrocytes to capillary blood vessels. Hiding like this, the infected cells try to avoid elimination in the spleen. Such massive accumulation of infected red blood cells in the capillary blood vessels of the brain and kidneys can lead to organ failure and ultimate death. Thus, the targeting of the parasite's RBC membrane proteins could aid the development of interventions that block the parasite's growth or limit the severity of the disease.

As reported in the PlasmoDB [31] database, there are 195 RBC membrane proteins containing HT motif and 293 RBC membrane proteins containing Pexel. In our study, we predict functions for 20 RBC membrane proteins containing either of the motifs. The list of RBC membrane proteins (with their predicted GO functions) containing both motifs is shown in Table 2 and the list containing one of the motifs is shown in Table 3. Some interesting examples, which could become future pharmaceutical targets, include RBC membrane proteins PFD0495c and PFE0040c assigned with gene ontology term GO:0007155 (cell adhesion) with probability 70% and 99% respectively. Furthermore,

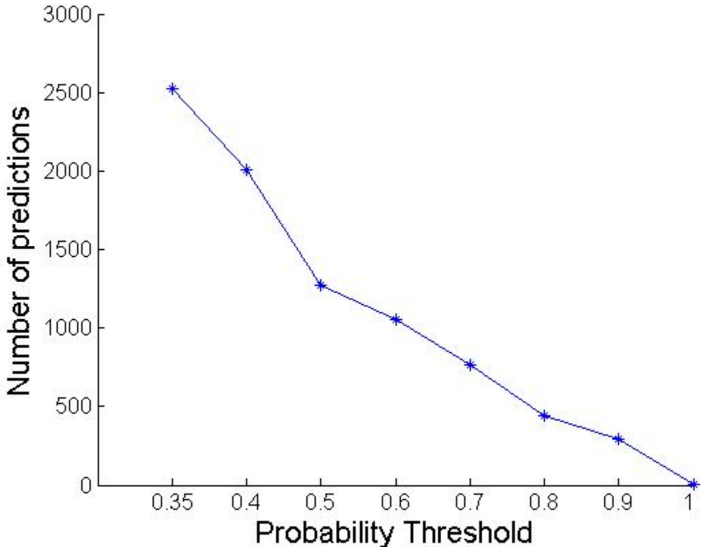

Figure 8: Number of possible predictions as a function of probability threshold. Each point corresponds to the number of predicted functional assignments whose probability is greater or equal to the corresponding probability threshold.

a special attention should be paid to gene ontology terms responsible for reaction to outside stimulus, as those can play a crucial role in the parasite's survival. For example, RBC membrane protein PFE1605w, assigned GO terms GO:0009628 (response to abiotic stimulus) with probability 80% and GO:0042221 (response to chemical stimulus) with probability 68%, could be a promising drug target.

Finally, there exist 10 *Plasmodium falciparum* surface proteins responsible for binding of the parasite to erythrocyte surface ligands, as reported by [31]. Following the establishment of a tight interaction between the parasite and the RBC, entry is initiated by the activation of actin-myosin motor so that the parasite forces the invagination of the erythrocytic membrane with formation of the the parasitophorous vacuole membrane, described earlier. The only surface Plasmodium protein, PFE0340c, present in our predicted dataset, is assigned GO terms GO:0006511 (ubiquitin-dependent protein catabolism) and GO:0019941 (modification-dependent protein catabolism) both with probability close to 63%.

5 Discussion and conclusions

In this paper, we have applied and evaluated a probabilistic approach for predicting protein functions for the malaria parasite *Plasmodium falciparum*. We combined four sources of information using a unified probabilistic framework. PPI and sequence

Protein ID	Probability	GO term
PFD0070c	0.401545231581066	GO:0043412 biopolymer modification
PFD0125c	0.49387370405278	GO:0006412 protein biosynthesis
	0.494801512287335	GO:0009059 macromolecule biosynthesis
	0.513352425272955	GO:0009058 biosynthesis
	0.512411033603626	GO:0044249 cellular biosynthesis
PFD0495c	0.580096975765904	GO:0006412 protein biosynthesis
	0.581846879650176	GO:0009059 macromolecule biosynthesis
	0.603310008429891	GO:0009058 biosynthesis
	0.602215349980034	GO:0044249 cellular biosynthesis
	0.699684816632941	GO:0007155 cell adhesion
PFD1020c	0.7280979853935	GO:0006631 fatty acid metabolism
PFD1170c	0.422344888143171	GO:0044267 cellular protein metabolism
	0.423250325426377	GO:0044260 cellular macromolecule metabolism
	0.432582635313454	GO:0019538 protein metabolism
	0.999999991137921	GO:0006457 protein folding
PFE0040c	0.98843772424871	GO:0007155 cell adhesion
PFE0060w	0.457539670421109	GO:0006468 protein amino acid phosphorylation
	0.521369701293125	GO:0006796 phosphate metabolism
	0.521369701293125	GO:0006793 phosphorus metabolism
	0.437130777000256	GO:0016310 phosphorylation
MAL7P1.170	0.476701149188691	GO:0006810 transport
	0.477943559067398	GO:0051234 establishment of localization
	0.477943559067398	GO:0051179 localization
PF07_0132	0.351905883602301	GO:0019538 protein metabolism
PFI1785w	0.365533239904503	GO:0019538 protein metabolism
	0.999794367636718	GO:0006457 protein folding
PF11_0508	0.375957294327154	GO:0006464 protein modification
PF13_0073	0.739599615802019	GO:0006412 protein biosynthesis
	0.733078093566159	GO:0009059 macromolecule biosynthesis
	0.751850612028445	GO:0009058 biosynthesis
	0.756174623097316	GO:0044249 cellular biosynthesis
PF13_0076	0.358502146993282	GO:0006810 transport
	0.360756570068609	GO:0051234 establishment of localization
	0.360756570068609	GO:0051179 localization
PF13_0275	0.358502146993282	GO:0006810 transport
	0.360756570068609	GO:0051234 establishment of localization
	0.360756570068609	GO:0051179 localization

Table 2: RBC membrane proteins possessing HT motif and Pexel, their predicted functions, and corresponding probabilities.

Protein ID	Probability	GO term
Pexel only:		
PFA0225w	0.487145531811038	GO:0043037 translation
	0.461767651699677	GO:0009058 biosynthesis
	0.453487695127242	GO:0044249 cellular biosynthesis
	0.539833659529562	GO:0006082 organic acid metabolism
	0.539833659529562	GO:0019752 carboxylic acid metabolism
	0.386239267057058	GO:0008610 lipid biosynthesis
	0.388721047331319	GO:0006629 lipid metabolism
	0.374685825510083	GO:0044255 cellular lipid metabolism
PFD0080c	0.665278152637513	GO:0006464 protein modification
	0.511094565590855	GO:0043412 biopolymer modification
	0.799046995682858	GO:0006468 protein amino acid phosphorylation
	0.684531881462785	GO:0006796 phosphate metabolism
	0.684531881462785	GO:0006793 phosphorus metabolism
	0.704629372061098	GO:0016310 phosphorylation
PFE1605w	0.802543920254657	GO:0044267 cellular protein metabolism
	0.754472317644877	GO:0044260 cellular macromolecule metabolism
	0.805311582644175	GO:0019538 protein metabolism
	0.72021237129596	GO:0006950 response to stress
	0.70563593694521	GO:0050896 response to stimulus
	0.801087730125495	GO:0009628 response to abiotic stimulus
	0.680269187401183	GO:0042221 response to chemical stimulus
	0.999999999999956	GO:0006457 protein folding
	0.501328353480413	GO:0007155 cell adhesion
MAL7P1.7	0.400917810998465	GO:0006082 organic acid metabolism
	0.400917810998465	GO:0019752 carboxylic acid
	0.999999999989668	GO:0006457 protein folding
PFI1780w	0.653500492148364	GO:0043037 translation
	0.877574807784855	GO:0006412 protein biosynthesis
	0.871609237465261	GO:0009059 macromolecule biosynthesis
	0.874807040307226	GO:0009058 biosynthesis
	0.356761397372017	GO:0044260 cellular macromolecule metabolism
HT motif only:		
PF13_0317	0.781092830960702	GO:0044267 cellular protein metabolism
	0.781906300484652	GO:0044260 cellular macromolecule metabolism
	0.78205791106515	GO:0019538 protein metabolism
	0.373719533733663	GO:0043037 translation
	0.781559322033898	GO:0006412 protein biosynthesis
	0.782192339038305	GO:0009059 macromolecule biosynthesis
	0.818767547332805	GO:0009058 biosynthesis
	0.818207742211449	GO:0044249 cellular biosynthesis

Table 3: RBC membrane proteins possessing only Pexel motif or only HT motif, their predicted functions, and corresponding probabilities

similarity data were presented in the form of functional linkage graphs, since such data imply the importance of the number and GO annotation of the nearest neighbors. Metabolic pathway and temporal gene expression data were encoded using categorical feature vectors, simplifying the search for similar feature patterns among related proteins.

We emphasized the importance of the data representation for parasites, though this might not necessarily apply to non-parasitic organisms. In particular, a malaria parasite's life cycle is affected by change of the host (e.g., mosquito and human), tissues (e.g., salivary glands, blood, gut wall, liver, red blood cells), and possible developmental changes of the parasite itself (e.g., gametocytes, sporozoites, merozoites). Each such change involves different mechanisms for gene regulation and employs many specific life-sustaining genes. Thus, it becomes crucial to analyze gene expression data from each stage separately, as opposed to calculating Pearson correlation coefficients for all pairs regardless of their temporal order. We have demonstrated that the data representation, which takes advantage of the temporal order of gene expression patterns, leads to a clear improvement in statistical significance over function predictions using simple Pearson coefficient calculations.

We show that data integration, previously shown to be beneficial for protein function prediction [5, 9, 10], is crucial when applied to organisms with limited individual data sources, as in the case of parasites. Even more importantly, the proposed "windowing" of the IDC provides a clear advantage to the data integration, dramatically improving its predictive performance. By embedding various data sources into the probabilistic framework, we have been able to assign functions to 628 previously unannotated *P. falciparum* proteins and expect to find in those some of the most promising candidates for future vaccine trials.

To extend this study to include ortholog genes, we next tested our method by integrating PPI data of another closely-related malaria parasite *P. vivax* (in particular, we used only PPI data of close orthologs with *P. falciparum*), and were encouraged by the significant improvement in the resulting performance scores and a much improved F1 curve. However, we have omitted further details of these improved results, since the *P. vivax* genomic data await publication and remain publicly unavailable. Once these data are published, we plan to disseminate the improved results through our laboratory website.

We believe that this work will pave the way for more complex automatic annotation algorithms based on model checking with temporal-logic queries—in this picture, one would obtain a succinct Kripke model (a phenomenological model) that summarizes the most important synchronization properties exhibited by a set of temporal data streams; then use these Kripke models to infer properties satisfied in various states (also called possible-worlds) of the model; and finally, associate these properties with functional classes and genes active in these states of the Kripke model. It should also be obvious that, at first, such a method is likely to be employed as a debugging tool for existing ontologies: particularly, to check if certain ontology terms are being associated incorrectly or inconsistently with a bio-molecule.

6 Acknowledgments

We would like to thank members of the NYU/Courant Bioinformatics group (particularly, Prof. Marco Antoniotti and Andrew Sundstrom) for many useful discussions, and Naoki Nariai of Boston University for his patience and help in answering many questions about the software usage and statistical analysis specifications.

References

[1] S. Letovsky and S. Kasif, “Predicting protein function from protein/protein interaction data: a probabilistic approach,” *Bioinformatics*, vol. 19, no. 1, pp. i197–i204, 2003.

[2] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, “Prediction of protein function using protein-protein interaction data,” *Journal of Computational Biology*, vol. 10, no. 6, pp. 947–960, 2003.

[3] M. Deng, T. Chen, and F. Sun, “An integrated probabilistic model for functional prediction of proteins,” in *Proceedings of the Seventh International Conference on Computational Molecular Biology (RECOMB)* (W. Miller, ed.), pp. 95–103, ACM, 2003.

[4] M. Deng, Z. Tu, F. Sun, and T. Chen, “Mapping gene ontology to proteins based on protein-protein interaction data,” *Bioinformatics*, vol. 20, no. 6, pp. 895–902, 2004.

[5] N. Nariai, E. Kolaczyk, and S. Kasif, “Probabilistic protein function prediction from heterogeneous genome-wide data,” *PLoS ONE*, vol. 2, no. 3, 2007.

[6] I. Lee, S. Date, A. Adai, and E. Marcotte, “A probabilistic functional network of yeast genes,” *Science*, vol. 306, no. 5701, pp. 1555–8, 2004.

[7] L. Lu, Y. Xia, A. Paccanaro, H. Yu, and M. Gerstein, “Assessing the limits of genomic data integration for predicting protein networks,” *Genome Research*, vol. 15, no. 7, pp. 945–53, 2005.

[8] O. Troyanskaya, K. Dolinski, A. Owen, R. Altman, and D. Botstein, “A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae),” *Proceedings of the National Academy of Sciences of the USA*, vol. 100, no. 14, pp. 8348–53, 2003.

[9] S. Carroll and V. Pavlovic, “Protein classification using probabilistic chain graphs and the gene ontology structure,” *Bioinformatics*, vol. 22, no. 15, pp. 1871–1878, 2006.

[10] A. Mitrofanova, V. Pavlovic, and B. Mishra, “Integrative protein function transfer using factor graphs and heterogeneous data sources,” in *IEEE International Conference on Bioinformatics and Biomedicine* (X.-W. Chen, X. Hu, and S. Kim, eds.), pp. 314–318, IEEE Computer Society, 2008.

[11] S. Altschul, T. Madden, A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[12] J. Liu and B. Rost, “Comparing function and structure between entire proteomes,” *Protein Science*, vol. 10, no. 10, pp. 1970–1979, 2001.

[13] J. Liu and B. Rost, “CHOP proteins into structural domain-like fragments,” *Proteins: Structure, Function, and Bioinformatics*, vol. 55, no. 3, pp. 678–688, 2004.

[14] J. Whisstock and A. Lesk, “Prediction of protein function from protein sequence and structure,” *Quarterly Review of Biophysics*, vol. 36, no. 3, pp. 307–340, 2003.

[15] M. Pruess, W. Fleischmann, A. Kanapin, Y. Karavidopoulou, P. Kersey, E. Kriventseva, V. Mittard, N. Mulder, I. Phan, F. Servant, and R. Apweiler, “The proteome analysis database: a tool for the in silico analysis of whole proteomes,” *Nucleic Acids Research*, vol. 31, no. 1, pp. 414–417, 2003.

[16] K. Nakai and P. Horton, “PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization,” *Trends in Biochemical Sciences*, vol. 24, no. 1, pp. 34–36, 1999.

[17] R. Nair, P. Carter, and B. Rost, “NLSdb: database of nuclear localization signals,” *Nucleic Acids Researh*, vol. 31, no. 1, pp. 397–399, 2003.

[18] B. Rost, J. Liu, R. Nair, K. Wrzeszczynski, and Y. Ofran, “Automatic prediction of protein function,” *Cellular and Molecular Life Sciences*, vol. 60, no. 12, pp. 2637–2650, 2003.

[19] A. Valencia and F. Pazos, “Computational methods for the prediction of protein interactions,” *Current Opinion in Structural Biology*, vol. 12, no. 3, pp. 368–373, 2002.

[20] M. Galperin and E. Koonin, “Who’s your neighbor? new computational approaches for functional genomics,” *Nature Biotechnology*, vol. 18, no. 6, pp. 609–613, 2000.

[21] A. Drawid and M. Gerstein, “A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome,” *Journal of Molecular Biology*, vol. 301, no. 4, pp. 1059–1075, 2000.

[22] U. Karaoz, T. Murali, S. Letovsky, Y. Zheng, C. Ding, C. Cantor, and S. Kasif, “Whole-genome annotation by using evidence integration in functional-linkage networks,” *Proceedings of the National Academy of Sciences of the USA*, vol. 101, no. 9, pp. 2888–2893, 2004.

[23] B. Schwikowski, P. Uetz, and S. Fields, “A network of protein-protein interactions in yeast,” *Nature Biotechnology*, vol. 18, no. 12, pp. 1257–1261, 2000.

[24] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh, "Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps," *Bioinformatics*, vol. 21, no. 1, pp. i302–i310, 2005.

[25] A. Butte and I. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pacific Symposium on Biocomputing* (R. Altman, A. Dunker, L. Hunter, and T. Klein, eds.), pp. 418–429, World Scientific, 2000.

[26] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the USA*, vol. 95, no. 25, pp. 14863–8, 1998.

[27] X. Zhou, M. Kao, and W. Wong, "Transitive functional annotation by shortest-path analysis of gene expression data," *Proceedings of the National Academy of Sciences of the USA*, vol. 99, no. 20, pp. 12783–8, 2002.

[28] S. Kleinberg and B. Mishra, "The Temporal Logic of Causal Structures," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, (Montreal, Quebec), *to appear*, June 2009.

[29] Z. Bozdech, M. Llins, B. Pulliam, E. Wong, J. Zhu, and J. DeRisi, "The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum," *PLoS Biology*, vol. 1, no. 1, 2003.

[30] D. LaCount, M. Vignali, R. Chettier, A. Phansalkar, R. Bell, J. Hesselberth, L. Schoenfeld, I. Ota, S. Sahasrabudhe, C. Kurschner, and et al, "A protein interaction network of the malaria parasite Plasmodium falciparum," *Nature*, vol. 438, no. 7064, pp. 103–107, 2005.

[31] A. Bahl, B. Brunk, R. L. Coppel, J. Crabtree, S. J. Diskin, M. J. Fraunholz, G. R. Grant, D. Gupta, R. L. Huestis, J. C. Kissinger, P. Labo, L. Li, S. K. McWeeney, A. J. Milgram, D. S. Roos, J. Schug, and C. J. Stoeckert, "Plasmodb: the plasmodium genome resource: An integrated database providing tools for accessing, analyzing and mapping expression and sequence data (both finished and unfinished)," *Nucleic Acids Research*, vol. 30, no. 1, pp. 87–90, 2002.

[32] S. Kleinberg, K. Casey, and B. Mishra, "Systems biology via redescription and ontologies (i): finding phase changes with applications to malaria temporal data," *Systems and Synthetic Biology*, vol. 1, no. 4, pp. 197–205, 2007.

[33] D. Altman and J. Bland, "Diagnostic tests. 1: Sensitivity and specificity," *British Medical Journal*, vol. 308, no. 6943, p. 1552, 1994.

[34] A. Maier, B. Cooke, A. Cowman, and L. Tilley, "Malaria parasite proteins that remodel the host erythrocyte," *Nature Reviews Microbiology*, vol. 7, pp. 341–354, May 2009.

[35] S. Reiff and B. Striepen, "Malaria: The gatekeeper revealed," *Nature*, vol. 459, pp. 918–919, June 2009.

[36] M. Marti, R. Good, M. Rug, E. Knuepfer, and A. Cowman, "Targeting malaria virulence and remodeling proteins to the host erythrocyte," *Science*, vol. 306, no. 5703, pp. 1930–1933, 2004.

[37] J. MacKenzie, N. Gomez, S. Bhattacharjee, S. Mann, and K. Haldar, "Functions in export during blood stage infection of the rodent malarial parasite plasmodium berghei," *PLoS ONE*, vol. 3, no. 6, p. e2405, 2008.