Systems Biology via Redescription and Ontologies (III): Protein Classification using Malaria Parasite's Temporal Transcriptomic Profiles

Antonina Mitrofanova
Computer Science department
New York University
antonina@cs.nyu.edu

Samantha Kleinberg
Computer Science department
New York University
samantha@cs.nyu.edu

Jane Carlton
Dept. of Medical Parasitology
New York University
Jane.Carlton@nyumc.org

Simon Kasif
Biomedical Engineering
Boston University
kasif@engc.bu.edu

Bud Mishra
Computer Science department
New York University
mishra@nyu.edu

Abstract

This paper addresses the protein classification problem, and explores how its accuracy can be improved by using information from time-course gene expression data. The methods are tested on data from the most deadly species of the parasite responsible for malaria infections, Plasmodium falciparum*. Even though a vaccination for Malaria infections has been under intense study for many years, more than half of Plasmodium proteins still remain uncharacterized and therefore are exempted from clinical trials. The task is further complicated by a rapid life cycle of the parasite, thus making precise targeting of the appropriate proteins for vaccination a technical challenge. We propose to integrate protein-protein interactions (PPIs), sequence similarity, metabolic pathway, and gene expression, to produce a suitable set of predicted protein functions for* P. falciparum*. Further, we treat gene expression data with respect to various changes that occur during the five phases of the intraerythrocytic developmental cycle (IDC) (as determined by our segmentation algorithm) of* P. falciparum *and show that this analysis yields a significantly improved protein function prediction, e.g., when compared to analysis based on Pearson correlation coefficients seen in the data. The algorithm is able to assign "meaningful" functions to 628 out of 1439 previously unannotated proteins, which are first-choice candidates for experimental vaccine research.*

1 Introduction

World-wide, each year, malaria infects approximately 515 million people and kills between one and three million of them. Better understanding of *protein functions* in the malaria parasite can be expected to produce a tremendous effect on approaches aimed at preventing current malaria epidemics. This anticipated impact is suggested by the fact that targets for drug design and vaccination are almost always based on proteins, especially those involving enzymatic functions. Unfortunately, since many *P. falciparum* proteins remain uncharacterized, they are mostly ignored by pharmaceutical laboratories and disregarded as potential protein targets in drug and vaccine development.

Toward this goal, the paper addresses the issue of automatic prediction of protein functions, using many sources of data, but with a particular emphasis on temporal transcriptomic profiles. 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. One promising computational approach utilizes probabilistic graphical models, such as belief networks, to infer functions over sets of partially annotated proteins [10, 6, 13]. Bayesian network methods for data integration have been extensively studied [16, 7, 12] to predict PPIs and protein function similarity for any pair of genes. Additionally, the approach of incorporating Gene Ontology dependencies into probabilistic graphical models [5, 15] 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 [1] analysis, and rely on the fact that similar proteins are likely to share common functions [11, 19, 17, 5]. At the same time, PPI data are widely used to infer protein func-

tions. For example, methods described in several recent papers [10, 6, 13] used the PPI 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 [10, 18, 3]. Clustering of genome-wide expression patterns has also been used to predict protein function, as described in [16, 14, 20].

In a majority of cases, *Saccharomyces cerevisiae* is chosen as a model, since it has been extensively studied from multi-omic view-points, and its protein data is also the most complete. The problem of protein function prediction is, however, more difficult in parasites (i.e. the malaria parasite), 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., its life-cycle of 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; etc). Such obstacles manifest themselves in a paucity of information on the protein properties, interactions, localization, motifs etc. of *Plasmodium* species. By relying on just one source of protein information, it is impossible to devise a reliable probabilistic framework with the ability to automatically predict classification for proteins of interest. 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.

Previously, it has proven beneficial to integrate heterogeneous data for predicting protein functions. Indeed, combining various types of information can improve the overall predictive power of automated protein/gene annotation systems for baking yeast, as shown in [16, 5, 15]. 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; and finally, gene ontology defines term-specific dependencies.

In our most recent work, we aimed to collect all information currently available for *P. falciparum* and to evaluate the predictive value of each source of data. 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 gene ontology classifications.

We stress the importance of the approach to the data used for protein function prediction in parasites. In particular, during *P. falciparum's* Intraerythrocytic Developmental Cycle(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 clustering time-course gene expression data from each stage of the cycle *separately* produces better results as compared with Pearson coefficient calculations applied to the time-course data as a whole.

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 even a few *P. falciparum* protein functions as these might play a significant role in the next stages of vaccine and drug development, leading to effective control of the disease.

2 Methods

2.1 Data

For our analysis, we focused on 2688 *P. falciparum* proteins from the time-course data [4], among which only 1249 proteins possess known biological process annotations. **Protein-protein interaction data:** We obtained Y2H (yeast-two-hybrid) data for *P. falciparum* from [9]. This dataset presents 1130 interactions covering 1312 proteins. **Sequence homology:** We started by gathering sequence information for proteins from [9]. Each sequence was queried against the entire *P. falciparum* sequence database [9] using BLAST. We recorded BLAST pairwise p-values 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. **Metabolic pathway data:** We used metabolic pathway data from [2]. The data consisted of 119 metabolic pathway categories for *P. falciparum*. The 3526 data pairs covered 1998 genes. **Temporal Gene expression data:** Time-course gene expression data covering the 48 hours of the Intraerythrocytic Developmental Cycle (IDC) of *P. falciparum* was obtained from a study by Bozdech et al. [4]. While the IDC comprised three main stages (Ring, Trophozoite, and Schizont, separated by two critical transition instants), the work in [8] identified four critical transition instants with major changes in gene regulation, corresponding to the five developmental periods ranging from 5 to 12 hours each. **GO data:** We used GO (gene ontology) terms as the basis of our annotation (in particular, 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 a protein, positively annotated to a GO term, 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. [16], we excluded labels that appear

less than five times among these genes, and defined 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 in two main 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 propose to build separate linkage graphs for PPI and sequence similarity, since, for these data, interacting partners are more likely to share a function. For PPI, the edges represent existing protein-protein interactions. For sequence similarity (homology), an edge is added when the pairwise p-value is less than 10^{-4}.

We adopted some ideas of the data representation and analysis of functional linkage graphs from Nariai et al. [16]. For each functional linkage graph l and for each Gene Ontology 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 performed 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.

A different method of data representation is the *categorical feature vector*, which holds a list of categories and assign 1 to a protein that belongs to a certain 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 it 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 ($r = 119$ is the number of metabolic pathway categories).

Finally, we propose to use a categorical feature vector, not the functional linkage graph, for gene expression profiles. Usually gene expression profiles are encoded into a functional linkage graph using the Pearson correlation coefficient calculated for all combinations of genes, as used in [16]. However, we believe that 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 [8] and applied k-mean clustering to the expression patterns of each time period, as described below. We considered proteins from the same cluster as those sharing the same categorical feature and thus possibly having 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 a protein where $d_r^j = 1$ if a protein is in cluster r in the time period j, and $d_r^j = 0$, otherwise. A feature vector then is $\mathbf{d} = (d_1^1, d_2^1, \ldots, d_q^1, d_1^2, d_2^2, \ldots, d_q^2, \ldots, d_1^w, d_2^w, \ldots, d_q^w)^T$, where $q = k$ is the number of clusters after k-mean clustering, and $w = 5$ is the number time windows.

For each protein i and each function t, we computed the posterior probability of this protein having a specific function. We adopt the basic ideas of such computation from [16] and present them in the extended version of the paper.

3 Experiments and Results

In 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 $Sensitivity = \frac{TP}{TP+FN}$ and $Specificity = \frac{TN}{FP+TN}$, where TP is the number of true positives, FN is the number of false negatives, etc. 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, etc. Each stage requires utilizing different life functions and possible metamorphosis, which up-regulates necessary genes and/or down-regulates 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. [8]. This expression

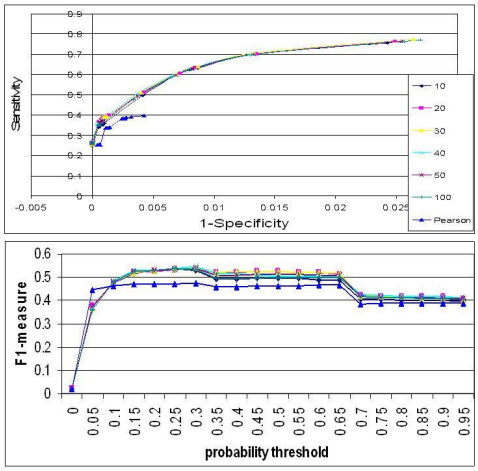


Figure 1. The ROC curve and F1 statistics of recall experiment by 5-fold cross validation for gene expression data. Numbered legends correspond to k-mean clustered datasets.

data is particularly interesting since the IDC, or blood stage, is the phase responsible for malaria symptoms in humans. This study [8] 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: $\mathcal{F}_{\min} = I(Z;X) + \beta\langle d(x,z)\rangle$, where mutual information and average distortion are: $I(Z;X) = \sum_{x,z} p(z|x)p(x)log\frac{p(z|x)}{p(z)}$, $\langle d(x,z)\rangle = \sum_{x,z} p(x)p(z|x)d(x,z)$, and $d(x,z) = \sum_{x_1} p(x_1|z)d(x_1,x)$.

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. Each window is then scored, based on its length and the above equations. 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. [8] found the critical time points at 7, 16, 28 and 43 hours, leading to 5 windows, sized rather 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.

In our method, we use these identified windows and cluster separately the expression profiles, delimited in each of them; for this purpose, we used k-mean 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 $\mathbf{d}$ of a protein.

We experimented with various values for k and compared results with the linkage graph defined by a Pearson coefficient calculation, as shown in Figure 1. In our experiments, due to a high number of negative annotations for the *P. facliparum* dataset, a ROC curve does not reflect a precise Sensitivity-Specificity relationship (since specificity reaches 0.9 immediately after threshold for posterior probability goes above 0.05) 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 $F1$ statistics, which reflects a relationship of Recall to Precision. Note that $F1$ will be maximized only if both measures are maximized.

As shown in Figure 1, 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 rage, the method yields nearly identical ROC and F1 curves. Thus, we fixed an arbitrary value, $k = 30$, for the following analysis.

Figure 1 also shows a clear superiority of time-dependent k-mean clustered data over Pearson coefficient dataset (in the majority of cases, Pearson curve is completely below the curves for the clustered data). The linkage graph defined by Pearson coefficient was built using 286620 edges (protein pair is considered co-expressed if Pearson coefficient is larger than 0.85 [16]) and covered 2646 proteins.

3.2 Analysis of Prediction Accuracy

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

Figure 2 summarizes the positive impact of data integration (PPI, sequence similarity, metabolic pathway, window-based gene expression clustering; gene expression by Pearson coefficients was not a part of the data integration) on protein function prediction via ROC and $F1$ measures. Since ROC curves are very much influenced by the large number of negative annotations in $P.\ falciparum$ data (similarly to Figure 1), the F1 statistics is more preferable.

Additionally, we investigated the impact of adding gene expressions to "fused" data (PPI, similarity, and metabolic pathway). In Figure 3, we show both ROC and F1 curves for fused data alone, then for fused data together with tem-

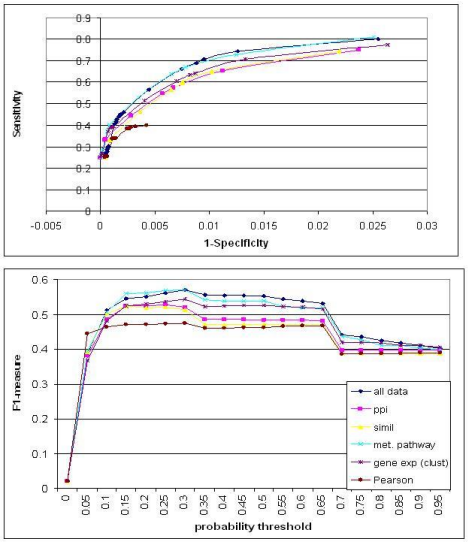


Figure 2. The ROC and F1 statistics for individual data sources and integrated data (posterior probability thresholds ranges from 0.05 to 0.95, in 0.05 increments).

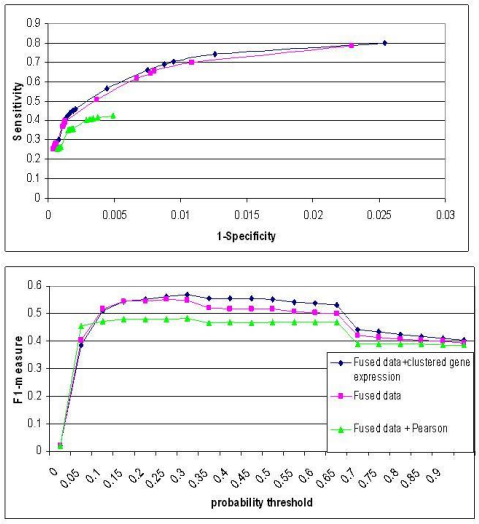


Figure 3. The ROC and F1 statistics for various ways of integrating data:"fused" is defined as ppi+similarity+metabolic pathway.

poral 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.

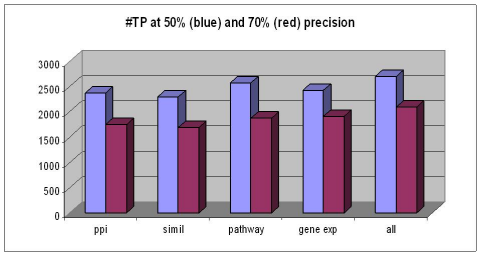
Figure 4 shows the impact of data integration on the number of TP at two precision levels: 50% and 70%. These two levels of precision are reasonably accurate of the range of possible improvements in our study, and the TP number is calculated when the precision level first hits the specified margin. As shown in the table of Figure 4, data integration significantly outperforms individual data sources at 70% precision, which corresponds to 0.35 threshold of posterior probability. 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 and tried to assign functions to proteins without annotations. 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 2546 gene-GO assignment pairs, which can be viewed at $www.cims.nyu.edu/\sim antonina/real\_output.txt$. The GO terms are reported together with their parents (ancestors) in the GO hierarchy.

4 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 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. At the same time, metabolic pathway and temporal gene expression data were encoded using categorical feature vectors, simplifying the search for similar feature patterns among related proteins.

We emphasize the importance of the data representation for parasites, though this might not necessarily apply to non-parasitic organisms. In particular, malaria parasites' 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, etc.), and possible developmental changes of the parasite itself (e.g., gametocytes, sporozoites, merozoites, etc.). Each such change involves different mechanisms of 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 calculation.



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

Figure 4. #TP at 50% precision (blue) and at 70% precision (dark red). Table describes % of improvements of data integration on #TP over individual data sources.

We show that data integration, previously shown to be beneficial for protein function prediction [16, 5, 15], is crucial when applied to organisms with limited individual data sources, as in the case of parasites. By embedding various data sources into the probabilistic framework, we are 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*, and were encouraged by the significant improvement in the resulting performance scores. Once the *P. vivax* genomic data are published, we plan to disseminate the improved results through our laboratory website.

More importantly, 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. 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. **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 her help in answering many questions about the software usage and analysis specifications.

References

[1] http://blast.ncbi.nlm.nih.gov/blast.cgi.

[2] http://plasmodb.org/plasmo/.

[3] S. B., U. P., and F. S. A network of protein-protein interactions in yeast. *Nat Biotechnol*, 18:1257–1261, 2000.

[4] Z. Bozdech, M. Llins, B. P. an ED Wong, J. Zhu, and J. DeRisi. The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol*, 1, 2003.

[5] S. Carroll and V. Pavlovic. Protein Classification Using Probabilistic Chain Graphs and the Gene Ontology Structure. *Bioinformatics*, pages 1871–1878, 2005.

[6] M. Deng, T. Chen, and F. Sun. An integrated probabilistic model for functional prediction of proteins. *RECOMB*, pages 95–103, 2003.

[7] L. I, D. SV, A. AT, and M. EM. A probabilistic functional network of yeast genes. *Science*, 306.

[8] 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*, 1(4), 2007.

[9] D. LaCount, M. Vignali, R. Chettier, A. Phansalkar, R. Bell, J. Hesselberth, L. Schoenfeld, I. Ota, S. Sahasrabudhe, C. Kurschner, et al. A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature*, 438(7064):103–107, 2005.

[10] S. Letovsky and S. Kasif. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19(1):197–204, 2003.

[11] J. Liu and B. Rost. Comparing function and structure between entire proteomes. *J.Mol.Biol*, 10:1970–1979, 2001.

[12] L. J. Lu, Y. Xia, A. Paccanaro, H. Yu, , and M. Gerstein. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res*, 15(7):945953, 2005.

[13] D. M., Z. Tu, F. Sun, and T. Chen. Mapping gene ontology to proteins based on protein-protein interaction data. *Bioinformatics*, 20(6):895–902, 2004.

[14] E. MB, S. PT, B. PO, and B. D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–8, 1998.

[15] A. Mitrofanova, V. Pavlovic, and B. Mishra. Integrative protein function transfer using factor graphs and heterogeneous data sources. *IEEE BIBM*, (to appear), 2008.

[16] N. N., K. E., and K. S. Probabilistic Protein Function Prediction from Heterogeneous Genome-Wide Data. *PLoS ONE*, 2(3):e337, 2007.

[17] M. Pruess, W. Fleischmann, A. Kanapin, Y. Karavidopoulou, P. Kersey, and et al. The Proteome Analysis database: a tool for the in silico analysis of whole proteomes. *Nucl. Acids Res*, 31(3):414–417, 2003.

[18] K. U., M. T., L. S., Z. Y., D. C., and et al. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci U S A*, 101:2888–2893, 2004.

[19] J. Whisstock and A. Lesk. Prediction of protein function from protein sequence and structure. *Quarterly Review of Biophysics*, 36:307–340, 2003.

[20] Z. X, K. MC, and W. WH. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci U S A*, 99(20):12783–8, 2002.