A data integration approach to predict host-pathogen protein-protein interactions: application to recognize protein interactions between human and a malarial parasite

Oruganty Krishnadev and Narayanaswamy Srinivasan*

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India

* Corresponding author

Email: ns@mbu.iisc.ernet.in
Phone: +91-80-2293 2837; Fax: +91-80-2360 0535

Edited by H. Michael; received January 15, 2008; revised April 06, 2008; accepted April 07, 2008; published May 17, 2008

Abstract

Lack of large-scale efforts aimed at recognizing interactions between host and pathogens limits our understanding of many diseases. We present a simple and generally applicable bioinformatics approach for the analysis of possible interactions between the proteins of a parasite, *Plasmodium falciparum*, and human host. In the first step, the physically compatible interactions between the parasite and human proteins are recognized using homology detection. This dataset of putative *in vitro* interactions is combined with large-scale datasets of expression and sub-cellular localization. This integrated approach reduces drastically the number of false positives and hence can be used for generating testable hypotheses. We could recognize known interactions previously suggested in the literature. We also propose new predictions which involve interactions of some of the parasite proteins of yet unknown function. The method described is generally applicable to any host-pathogen pair and can thus be of general value to studies of host-pathogen protein-protein interactions.

Keywords: host-pathogen interactions, protein-protein interactions, data integration method

---

Introduction

The biochemical function of a protein to a large extent is strongly correlated to the three dimensional structure it adopts. In an *in viv*o condition, apart from adopting a suitable conformation a protein can only function in a biologically meaningful manner if its function is regulated. This regulation is usually either at the level of gene expression or by interacting with other proteins in the cell [Deng *et al.*, 2004; Jenner and Young, 2005]. Most pathogens disrupt the delicate balance of the regulatory processes in the host cell by disrupting the processes at both these levels [Merrell and Falkow, 2004]. The disruption usually takes place due to the interplay of interactions between the host proteins and pathogen proteins. To a first approximation, the pathogenesis of a disease is an outcome of the interactions between host and pathogen proteins and other kinds of molecules and hence, detection of interactions between proteins of the host and pathogen organisms is essential for a complete understanding of the pathogenesis. The disruption of such protein - protein interactions by small molecules is a step towards discovery of drugs which are specific for the disease [Loregian and Palu, 2005; Zutshi *et al.*, 1998]. At present, drug discovery by inhibition of specific protein-protein interactions important for pathogenesis, is impeded due to the unavailability of datasets of interacting proteins across host and pathogen organisms.

The study of protein-protein interactions on a large scale was accelerated, after the genome sequences were made available, to gain a better understanding of the physiology of the organisms under study [Legrain *et al.*, 2001]. One of the early analyses was performed on yeast using the yeast two hybrid approach [Uetz *et al.*, 2000; Ito *et al.*, 2001]. Such analyses provided data on the protein-protein interactions which might be occurring in the yeast cell. The yeast two hybrid approach has been extended to many organisms and the interaction maps are available for *D. melanogaster* [Giot *et al.*, 2003], *C. elegans* [Li, S. *et al.*, 2004], *H. pylori* [Rain *et al.*, 2001] and a skeletal interaction map for the human genome [Stelzl *et al.*, 2005] is also available. There are still many organisms for which the genomes have been completely sequenced and yet the interactions between the proteins of the organism are not yet identified. This led many groups to extrapolate the interactions seen in one organism to another organism by homology-based information transfer [Wojcik *et al.*, 2003; Han *et al.*, 2004]. Such interactions which have been inferred by homology to a known interaction are known as 'interologs'. The study of interologs has traditionally been limited to interactions occurring between proteins of a single organism. Since homology alone cannot suffice to identify endogenously feasible interactions, information from a large number of datasets is integrated to get a comprehensive list of interacting proteins [Huang *et al.*, 2004; Persico *et al.*, 2005; Rhodes *et al.*, 2005; Zhong and Sternberg, 2006]. The data used are usually expression and/or localization information. The utility and the accuracy of homology-based methods have been studied extensively and it has been found that close homologues can be expected to have similar sub-cellular localization patterns [Nair and Rost, 2002] and similar enzymatic functions [Tian and Skolnick, 2003]. More importantly, distant homologues adopt similar three-dimensional structures as demonstrated in manually curated structural classification databases like SCOP [Murzin *et al.*, 1995]. It has also been demonstrated that close homologues interact structurally in similar manner [Aloy *et al.*, 2003]. Thus, homology based methods have been

shown to predict reliably at least two factors necessary for interaction of proteins *in vivo* namely spatial proximity and the ability to bind physically.

We suggest that using the 'interologs' approach, integrated data can be generated for host - pathogen interactions as, structurally, there is no difference between interactions between proteins of the same organism or from different organisms. In both cases, homology gives an indication of the ability to physically bind together or of functional interaction between proteins. Such an approach has been proposed by Dyer *et al.*, 2007, and the method suggested by them gives a small list of interactions which is not rigorously evaluated for its biological relevance. An exhaustive approach for enumerating the protein-protein interactions between host and pathogen organisms based on comparative modeling of 3-D structures has been suggested by Davis *et al.*, 2007. Since homology to known interactions is not sufficient for assessment of biological relevance, the localization and gene expression information should be used to assess if the interaction is possible under an *in vivo* condition. For instance, two proteins may interact *in vitro* but may not interact *in vivo* if the proteins are localized to different compartments of the cell. Thus, homology can give information about *in vitro* interactions. In the case of host-pathogen interactions, the task is made difficult as pathogens might be present in different compartments at different times of the pathogenesis. In this paper we integrate protein-protein interaction data with experimental and predicted information about human and *P. falciparum* proteins expression and localization.

We chose *P. falciparum* as it causes malaria, one of the most prevalent diseases in third world countries, and also because the genomes of the pathogen *P. falciparum*, and the host H. sapiens have been sequenced [Hoffman *et al.*, 2002; Ridley, 2002] aiding large-scale computational efforts to interrogate host-pathogen interactions through bioinformatic approaches. Another advantage of using *P. falciparum*-human interactions is that the pathogen is known to invade and reside inside different cell types of the host at different points in the life cycle of the pathogen. Thus, the stage-specific expression patterns are clearly demarcated in case of *P. falciparum* and can be very well correlated to organ-specific expression data in human host. Both organisms have been studied extensively after the genomes have been sequenced and the yeast two-hybrid like approach has been applied to protein-protein interaction network of the pathogen and thus many of the interactions between *P. falciparum* proteins are known experimentally. The disparate data sets on the expression and localization can be integrated to gain an understanding of the pathogenesis of malaria. This approach of data integration to study host-pathogen protein-protein interactions is not restricted only to the combination of *P. falciparum* and human and can be applied to any pathogen-host pair. The context of the interactions in different pathogens can differ, and thus only the assessment of the interactions would differ in different organisms. Hence the method presented here can be applied to any host-pathogen pair.

Methods - A new approach to study host-pathogen protein-protein interactions

Generation of interaction dataset

The first step in the generation of the interaction dataset uses homology detection methods and requires a dataset of known protein interactions. We chose to employ two different datasets to get a comprehensive list of putative interactions. One of these databases is the DIP (Database of Interacting Proteins) [Salwinski *et al.*, 2004] which is curated semi-automatically to provide a set of interactions which are deemed to be physically interacting. The list of DIP interactions has been enriched with the interactions among *P. falciparum* proteins identified using the two-hybrid approach in *P. falciparum* [LaCount *et al.*, 2005]. The other dataset we used is the iPfam [Finn *et al.*, 2005] dataset which provides the interactions occurring between functional domains present in Pfam [Finn *et al.*, 2006]. The interaction between two functional domains is suggested in iPfam based on the presence of the homologues in a complex of known three dimensional structure. Thus, the iPfam dataset provides a resource for the detection of structurally compatible interactions.

The sequences of the proteins present in enriched DIP dataset were used to query the protein sequences of *P. falciparum* obtained from PlasmoDB [Bahl *et al.*, 2003] and human protein sequences obtained from the NCBI site (ftp://ftp.ncbi.nlm.nih.gov/genomes/human/). PSI-BLAST [Altschul *et al.*, 1997] has been used for sequence searches with the enriched DIP database with *E*- and *H*-value cut-offs of 10^{-5} and 10^{-6} respectively. In addition to the cut-offs used in the PSI-BLAST program, we also analyzed the initial hits for coverage on the query sequences (DIP sequences) in the alignment. The coverage was kept at 80% with the identity between query and hit sequence being greater than 30%. The coverage value was chosen as it minimizes the occurrence of false positives and was found out by querying the homologues of complexes with known structure and analyzing the feasibility of interaction by evaluating the conservation of interfacial residues (data not shown).

The functional domains were assigned using the MulPSSM approach previously developed in our group. The MulPSSM [Gowri *et al.*, 2006] approach enables a sensitive search to be performed, using RPS-BLAST, on a dataset of family profiles in which every protein domain family is represented by multiple position-specific scoring matrices. The *E*-value cutoff for the domain assignments was fixed at 10^{-6} with the coverage on the Pfam family profile fixed at 60% to minimize the occurrence of false positives.

The PSI-BLAST search of DIP database sequences against the host and pathogen genomes and the interactions obtained based on the iPfam approach results in a preliminary dataset of protein-protein interactions occurring across genomes as well as within the genomes. From this gross list biologically viable interactions have been identified as described in the next section. A flowchart of the steps involved in the generation and analysis of the dataset is given in Fig. 1.

Click on the thumbnail to enlarge the picture

Figure 1: Flowchart of the analysis. Schematic diagram of the major steps followed for the analysis of interactions between human and *P. falciparum* proteins. The numbers of interactions filtered at various stages of the analysis are also indicated.

Steps of pruning in the identification of protein-protein interactions

Removal of ubiquitous interactions

Many of the protein-protein interactions across host and guest, identified in the preliminary list are conserved within almost all organisms studied (e. g. cytochrome b - cytochrome c interaction). To remove these spurious cases, the DIP sequences were queried against completely sequenced genomes of organisms belonging to various phyla. These organisms were *A .thaliana*, *D. melanogaster*, *A. gambiae*, *S. cerevisiae*, *E. coli* and *C. elegans*. Any interaction present within each of the above organisms is suggested to be less likely to happen between a host and a pathogen. Hence such interactions between host and pathogen are given less importance in the subsequent analysis but were not removed from the consolidated list. The interaction dataset was queried for proteins such as ubiquitin, DNA polymerase, splicing factors and such ubiquitous proteins which may not be expected to interact. All the interactions which involved such proteins were removed from the analysis.

Assessment of interactions

Since the protein-protein interaction template dataset used consists of interactions from many organisms which are evolutionarily distant from *P. falciparum* or human, it is possible that many interactions detected would not be plausible due to different localization of the proteins in the cells. Thus, the sub-cellular localization of the proteins was taken into consideration. The predicted sub-cellular localization information for human proteins was taken from H-ANGEL [Imanishi *et al.*, 2004] database which comprises of predictions from PSORT [Gardy *et al.*, 2005], TMHMM [Krogh *et al.*, 2001], SOSUI [Hirokawa *et al.*, 1998] and TargetP [Emanuelsson *et al.*, 2000]. For *P. falciparum* proteins, the localization data was not available from a single source. We predicted the localization of *P. falciparum* proteins using TMHMM (default values) and TargetP (default values). For TargetP interactions, two predictions were made: using the plant-type and non-plant type option given by TargetP. Any protein for which the localization information is different in plant-type option and non-plant type option was given a tag of undefined location. Another dataset which predicts the 'exportome' of the parasite proteome was used to identify the proteins which have a host targeting signal [Sargeant *et al.*, 2006].

Inclusion of expression data to enhance the information content

Tissue specific expression data from the tissue atlas of Genomics Initiative of Novartis Foundation [Su *et al.*, 2004] was taken and the information about expression of proteins in the liver and blood was

extracted from the data. The expression data available in the Unigene database [Wheeler *et al.*, 2003] was also used after extracting the liver and blood specific expression information of the proteins of human proteome. The expression profile of the *P. falciparum* proteins was taken from microarray and mass spectrometric data provided in PlasmoDB database and was divided into 7 stages viz sporozoite, merozoite, schizont, ring, trophozoite, gametocytes and gametes. Since *P. falciparum* infects different organs at different stages, functional context can be assessed using the expression and localization information. The expression data was taken from the published databases [Ben Mamoun *et al.*, 2001; Florens *et al.*, 2002; Lasonder *et al.*, 2002; Le Roch *et al.*, 2002; Bozdech *et al.*, 2003]. As all the experiments do not cover all the proteins and all the stages, we deemed a protein to be expressed in a stage if at least one of these studies suggests that it is expressed in that stage.

If the expression information and sub-cellular information was consistent with the interaction of the proteins from the putative interaction list obtained, it is considered as a structurally and functionally compatible interaction. All interactions present in the dataset have been analyzed manually, in the light of the available literature information, to assess the biological significance of the interaction.

Results

The data integration approach proposed in this paper has been applied to study *P. falciparum* and human protein-protein interactions. We define the interaction between the two organisms as a host-guest organism pair wherein one organism, the guest, resides inside the host organism for a part of its life-cycle. In our analysis human cells are the host and *P. falciparum* is the guest which resides inside the human host during a part of the life cycle of the parasite. In such a pair it is inevitable that the guest organism will have extensive interactions with the host for ensuring its survival by immune evasion, and hijacking the host machinery for its reproduction. It is reasonable to assume that many of the interactions between the host-guest pair will be mediated through proteins and these are the kind of interactions which are studied in this work. In the next section, the rationale for the work and results obtained from the work are given. In the section following it, some of the biologically significant examples coming out of this analysis will be discussed.

Generation of interacting pairs of host-pathogen proteins

The generation of structurally interacting protein pairs has been pursued with the supposition that closely related homologous proteins would interact in the same manner [Rekha *et al.*, 2005] in different organisms. This conclusion has been demonstrated in many cases and different groups differ on the criteria employed to detect close homologues. In our work, we have used highly stringent criteria (details given in Methods section) that minimizes the possibility of occurrence of false positives. Proteins from the host and pathogen genome are queried using profile based homology search methods

and a putative interaction set is created as a first step. This first list of interactions between the host and pathogen proteins is then refined using the expression and sub-cellular localization information to get a refined set of interaction which have a higher probability of interaction. An overview of the method followed in the analysis is given as a flowchart in Fig. 1.

Homologues of protein pairs interacting in one organism need not interact in another organism as the context in which they interact (for e. g., same sub-cellular localization) may not be present in a different organism. Thus in general, it is necessary to establish that the context of the template interactions exists in another organism also. In the present work, it is difficult to assess the interactions as these are not occurring within an organism but across two organisms and thus many proteins would not be physically close enough to interact. We overcame this difficulty by recognizing that *P. falciparum* interacts with different cell types of human organs at different stages of its life-cycle. We had considered the cell type-specificity of human proteins with physical compatibility to interact with *P. falciparum* proteins. For example, in some stages, *P. falciparum* is observed to be present residing inside the host cells and so the interactions in such stages would be different from the interactions in stages where invasion is about to take place. For example, *P. falciparum* is injected into the human host in the sporozoite stage and upon entry specifically adheres to liver cells and invades them [Kappe *et al.*, 2003]. Thus, there is high likelihood that extensive interactions between liver specific membrane proteins and *P. falciparum* membrane proteins expressed in sporozoite stage leads to invasion by the pathogen. In the Schizont stage on the other hand, a stage where multiple copies of the pathogen break open the host cell and are released into the blood stream, there is higher likelihood of interactions between membrane or secreted proteins of *P. falciparum* interacting with intracellular proteins of human host. Thus, the information needed to assess the context of interaction falls into two categories: 1) Tissue-specific expression information in the human host and stage-specific expression of *P. falciparum* proteins. 2) Spatial information comprising of subcellular localization information about human proteins and information about membrane or extra cellular nature of proteins in *P. falciparum*. It should be noted that the expression and spatial information is required to assess interactions between any host-pathogen pair and thus the method described here can be applied to any host-pathogen pair.

Details of the number of proteins present in *P. falciparum* and human and the number of interactions at various stages of filtering are shown in Tab. 1. Out of a large number of putative interactions, there were 43346 interactions (between 1002 *P. falciparum* proteins and 13660 human proteins) for which the expression information was found to be compatible. The full list of interactions is given as supplementary material (Supplementary Table 1 and Supplementary Table 2) where the interactions listed have not been pruned using sub-cellular localization information as the localization information is not from experiments and hence can lead to rejecting interactions which might be occurring *in vivo*.

Table 1: Distribution of *P. falciparum* proteins into stages which affect human host.

Stage	Total number of proteins	Number of interactions with compatible expression	Number of interactions with compatible expression and sub-

			cellular localization
Sporozoite	2147	27884	85
Merozoite	2115	26121	116
Ring	2866	30527	125
Trophozoite	3493	33074	157
Schizont	3722	22032	1226

The total number of proteins and the number of proteins found to interact with human proteins are given in the table. The numbers in the last column are for those proteins where sub-cellular localization information could be reliably associated. There are many cases in which sub-cellular localization is not clearly assigned and these cases are not included in the count.

In the next few sections, the details of interaction in some important stages of pathogenesis viz. sporozoite, merozoite and intra-erythrocytic stages are given in brief demonstrating that the data integration method can remarkably reduce the number of putative false positives. The total number of interaction predicted for proteins expressed in each of the stages of the lifecycle of *P. falciparum* is listed in Tab. 1. The table also lists interactions for which the localization information was also found to be compatible with an *in vivo* interaction.

Examples of biologically significant interactions

The assessment of biological significance of predicted interactions can be made by considering manually the available literature information on the proteins concerned. Hence the examples given in this section are not intended to be comprehensive and it is possible that more significant examples might be present in our dataset of predicted interactions. The list of a few examples of interest is given in Tab. 2 and some of the examples are discussed. It must be pointed out that not all the proteins of the pathogen which are known to be important for pathogenesis have been identified as a part of our predicted list of host-pathogen interactions. This is due to a critical step employed in our approach of finding similarity to a known interaction. The proteins in the host or pathogen proteome that are unique to the organism or have diverged considerably would not be recognized as similar to any protein in the template interaction datasets and hence *P. falciparum*-specific proteins involved in interaction with human proteins are unlikely to figure in our list of predicted interactions. Nevertheless the proposed method is capable of recognizing biologically feasible interactions not known before.

Table 2: *Plasmodium falciparum* proteins which are important for pathogenesis and their predicted interactions with human proteins.

Protein name	Annotation	Predicted membrane	Expressed in stages [a]	Human proteins predicted to interact

		region/signal peptide/export signal		
PFC0640w	CSP and TRAP-related protein (CTRP)	signal peptide	SP, RI, TZ, SZ, MZ, GC	1. ICAM-1 2. ICAM-4
PF13_0201	TRAP/SSP2 protein	none	SP, RI, TZ, SZ, MZ, GC	1. ICAM-1 2. ICAM-4
PF14_0075	plasmepsin, putative	membrane region	SP, RI, TZ, SZ, MZ, GC, GA	1. GTP binding protein 2. moesin 3. ezrin 4. radixin
PF14_0076	plasmepsin 1 precursor	membrane region	RI, TZ, SZ, MZ, GA	1. GTP binding protein 1 2. moesin 3. ezrin 4. radixin
PF14_0077	plasmepsin 2	membrane region	SP, RI, TZ, SZ, MZ, GC, GA	1. GTP binding protein 1 2. moesin 3. ezrin 4. radixin
PFC0495c	aspartyl protease, putative	none	SP, SZ, GC	1. GTP binding protein 1 2. moesin 3. ezrin 4. radixin
PFI1475w	merozoite surface protein 1	membrane region	SP, RI, TZ, SZ, MZ, GC	1. coagulation factor III precursor 2. EGF-containing fibulin-like extracellular matrix protein 1 precursor 3. complement

				component 9 4. integrin, beta 2 precursor 5. C-type lectin domain family 4, member E
PFE0120c	merozoite surface protein 8	membrane region	RI	1. coagulation factor III precursor 2. EGF-containing fibulin-like extracellular matrix protein 1 precursor 3. complement component 9 4. integrin, beta 2 precursor 5. C-type lectin domain family 4, member E
PF14_0635	hypothetical protein, conserved	none	RI, TZ	1. similar to alpha-actinin 4 2. similar to visinin-like protein 2 3. myosin light chain 1 4. guanylate cyclase activator 1B 5. centrin 1 6. alpha-spectrin (fragment)
PFI0095c	hypothetical protein	export signal	SP, RI, TZ, SZ, GC,	1. ankyrin repeat domain 28 2. vascular cell adhesion molecule 1 isoform a precursor
PFC0140c	N-ethylmaleimide-sensitive fusion protein, putative	none	SP, RI, TZ, SZ, MZ, GC	1. DsRBD domain containing protein 2. DEAD/H (Asp-Glu-

				Ala-Asp/His) box polypeptide 32 3. DEAD/DEAH box helicase domain containing protein
PFE0445c	SNAP protein (soluble N-ethylmaleimide-sensitive factor attachment protein), putative	none	RI, TZ, SZ, GC	1. Sec1-like protein family protein 2. syntaxin 4 3. U5 snRNP- specific protein 4. carnitine/acylcarnitine translocase 5. EPIM protein 6. ADP,ATP carrier protein, heart/skeletal muscle isoform T1

The stages in which the *P. falciparum* protein is expressed and presence of membrane spanning regions/signal peptides is also given in the table. Only a few representative interactions which were found to be biologically significant have been reported in this table.

[a] The abbreviations used for the stages are as follows: SP = sporozoite, MZ = merozoite, RI = ring, TZ = trophozoite, SZ= schizont, GC = gametocytes, GA = gametes

The first four proteins listed in Tab. 2 are the aspartic proteases plasmepsins (PF14_0075, PF14_0076 and PF14_0075) and a putative aspartyl protease (PFC0495w). *P. falciparum* contains many plasmepsins which play a crucial role in hemoglobin degradation and inhibiting these proteases affects the growth of the parasite. Another role suggested for the plasmepsins in literature is degradation of host cytoskeletal protein spectrin [Le Bonniec *et al.*, 1999]. It has been shown for example, that addition of protease inhibitors to a culture of *P. falciparum* inhibits the release of pathogens from the erythrocytes [Wickham *et al.*, 2003]. This suggests a role for proteases in host cytoskeleton degradation and subsequent release of pathogens into the bloodstream. Plasmepsin II has been shown to localize to the erythrocyte membrane and also cleave a cytoskeletal protein spectrin *in vitro* [Le Bonniec *et al.*, 1999]. From our analysis, we find that the *P. falciparum* plasmepsins interact with the cytoskeletal proteins of the host. These interactions are between the aspartyl protease-like region on the plasmepsin and host moesin, ezrin and G βsubunit. The host proteins moesin and ezrin are regulatory proteins which maintain the integrity of the cytoskeletal network [Mangeat *et al.*, 1999]. Disruptions in moesin and ezrin function could lead to destabilization of the erythrocyte membrane. Even more

remarkable is the fact that even though *P. falciparum* might regulate its cytoskeletal network through similar mechanisms, the plasmepsins in our predictions do not interact with any of the other *P. falciparum* cytoskeletal proteins. The predicted interactions of the plasmepsins are schematically depicted in Fig. 2a. This interaction has not been suggested before and the biological context is very appropriately poised to enable our approach to predict this interaction. The DIP interaction which enabled our approach to predict this interaction is between erm-1 and aspartyl protease of *C. elegans* that was identified using genome-wide two-hybrid approach in *C. elegans*.

a)

Click on the thumbnail to enlarge the picture

b)

Click on the thumbnail to enlarge the picture

Figure 2: Schematic diagram of the interactions. Rectangles indicate *P. falciparum* proteins and ellipses indicate human proteins. All the images were drawn using the software GraphViz [Gansner, 1999].

(a) Interactions predicted to occur between *P. falciparum* plasmepsins and other *P. falciparum* and human proteins. The interactions in this example are between *P. falciparum* proteins and human proteins with no homologous interaction present between *P. falciparum* proteins.

(b) Interactions between a hypothetical protein PF14_0635 and other *P. falciparum* and human proteins. The interactions in this example are of two kinds. In the first type, the *P. falciparum* protein given in the centre is predicted to interact with both human proteins and homologous *P. falciparum* proteins (shown in dashed lines). In the second type, the *P. falciparum* protein interacts only with *P. falciparum* proteins with no human protein homologue or only with human proteins with no *P. falciparum* homologue.

Our approach has detected some interactions between host and pathogen proteins that were previously suggested in literature to interact. Such observations provide strong support to the proposed approach to predict host-pathogen protein-protein interactions. Most important among these 'predictions' of known interactions is the one between Sporozoite surface protein 2 (SSP2 - PF13_0201) and various host ligands. The SSP2 protein also known as thrombospondin related anonymous protein (TRAP) contains an A domain which is known to mediate interactions with collagen [Ueda *et al.*, 1994], intercellular adhesion molecule 1 (ICAM-1) [Randi and Hogg 1994], and coagulation factors [Altieri and Edgington, 1988]. Targeted disruptions in the SSP2/TRAP protein have shown that this protein is essential for the invasion of mammalian liver cells [Sultan *et al.*, 1997]. The interactions predicted by our method which involve this important protein of the pathogen are summarized in Tab. 2. Another protein which has an A domain and according to the microarray data used in this work is expressed in various stages of the life cycle of *P. falciparum* is CTRP (CSP and TRAP related protein). Our method predicts interactions of CTRP with various host ligands which have compatible expression and sub-cellular localization. The CTRP protein is known to be important for the invasion of ookinetes in the mosquito mid-gut [Dessens *et al.*, 1999]. Our analysis and the available data suggest that this protein can play a role in the invasion process occurring in the human tissues. Thus, the proposed approach is capable of generating new hypothesis on the pathogenesis of the disease which can be explored experimentally.

Another group of proteins which have been implicated in the invasion of RBC is the merozoite surface proteins [Singh *et al.*, 2003]. We have detected interactions of MSP-1 and MSP-8 with various host receptors. The MSP-1 protein has been suggested to interact with band 3 protein of RBC's [Goel *et al.*, 2003; Li, X. *et al.*, 2004]. Our approach has not detected this particular interaction. We suggest alternative ligands for the MSP proteins based on homology to known protein-protein interactions. Experimental exploration of such alternatives will provide new hypothesis regarding the pathogenesis of *P. falciparum* malaria.

Most of the clinical symptoms are manifested in the intra-erythrocytic stages of the malarial life-cycle. Since RBCs do not have much of the cellular machinery usually present in eukaryotic cells, many proteins of the pathogen are exported to the RBC in order to acquire nutrients, remodel the host membrane for schizogony and cellular trafficking of exported proteins. Experimental studies have suggested two motifs which are thought to be necessary and sufficient for export to the RBC [Hiller *et al.*, 2004; Marti *et al.*, 2004]. Recently, using bioinformatics tools the complete 'exportome' of the parasite has been deduced [Sargeant *et al.*, 2006]. Our method predicts interactions for many of the proteins suggested to be exported to the RBC. These include a few protein kinases, falcipain and many hypothetical proteins. It is interesting that our method has resulted in prediction of interactions for hypothetical proteins in *P. falciparum*. Some of the host proteins which are suggested by our method to interact with *P. falciparum* hypothetical proteins include ankyrin repeat containing proteins. The ankyrin repeat proteins are present in the cytoskeletal network and disruption of such interactions can facilitate the rupture of host membrane which would release the mature parasites into circulation.

The role played by *P. falciparum* proteins in the trafficking of the parasite proteins is currently not

known very well. In this context, two proteins listed in Tab. 2, namely PFC0140c (N-ethylmaleimide-sensitive fusion protein, putative) and PFE0445c (soluble N-ethylmaleimide-sensitive factor attachment protein, putative) are significant. These two proteins have been suggested in the literature to be involved in subverting host cellular trafficking mechanisms to export *P. falciparum* proteins [Taraschi *et al.*, 2003] rendering support to our prediction.

Another conserved hypothetical protein of unknown function (PF14_0635) predicted to interact with human proteins is expressed in the ring and trophozoite stage only according to the microarray datasets used in this analysis. PF14_0635 interacts with human actin, myosin, centrin, calmodulin and spectrin proteins. PF14_0635 also interacts with *P. falciparum* centrin, and calmodulin. The interaction with actin and spectrin, however, is seen only between the human and *P. falciparum* proteins. The interactions of this protein with human and *P. falciparum* proteins are shown schematically in Fig. 2b. This protein might have been conserved due to its role in cytoskeletal regulation.

Discussion

We have presented the analysis of prediction of protein-protein interactions occurring across organisms with host-guest relationship. The analysis was carried out on *P. falciparum*-human interactions but the method can easily be extended to other host -pathogen organisms. Our work of extrapolation of interologs to cross-genome interactions is according to our knowledge the first such approach. Thus, the extent to which this method can be applied for detection of host-pathogen interactions needs to be further analyzed to test the reliability of the predicted interactions. However, such an assessment is difficult to make because of lack of any experimentally curated large-scale database of interactions between any two host and pathogen genomes. At present, no comprehensive experimental results seem to be available on protein-protein interactions between *P. falciparum* and human. Hence the interactions detected by us form the first attempt in this important direction. While the proposed data integration approach is not expected to be capable of identifying interactions involving proteins unique to the organisms involved, it can nevertheless identify other biologically viable interactions occurring across the host and pathogen. It seems that the approaches such as the one proposed in this paper are currently the best strategy for such genome-wide studies due to the absence of experimental protocols aimed at genome-wide detection of host-pathogen protein-protein interactions.

The remarkable result which has been shown in the examples covered in the present work is that even with limited data, testable hypotheses about the interaction which might be important for pathogenesis can be made. For example, the interactions between plasmepsins and host cytoskeletal proteins, interaction between TRAP and ICAMs suggested in the literature are also predicted in our analysis. Our method also predicts novel interactions which could lead to new directions in the experimental exploration of pathogenesis. Another interesting feature of our analysis is the prediction of interactions for hypothetical proteins which paves the way for new hypothesis regarding pathogenesis to be

generated. The fact that such interactions are picked in our analysis suggests that data integration methods like ours can provide clues to the pathogenesis of different organisms and the accuracy of such prediction can only increase with the availability of more experimental data. For example, *ab initio* approaches for prediction of protein-protein interactions can lead to further enhancement in the coverage of the host-pathogen protein-protein interactions (see for example Shen *et al.*, 2007).

The method proposed by us has a critical step of detecting homology to a known interaction. This places a limit on the number of interactions which can be predicted using our method. Many of the proteins in an obligate parasite such as *P. falciparum* do not show any similarity to proteins of other organisms. Thus it is difficult to predict interactions in such cases. *Ab initio* approach for predicting protein-protein interactions proposed by Shen *et al.*, 2007, could be useful in such situations. The accuracy of prediction is further complicated by the inaccuracies inherent in the datasets of known protein-protein interactions which mainly lists genome-wide yeast two hybrid data. In the case of DIP interactions, curation partially reduces the rate of false positives but does not completely remove it. The interactions suggested on the basis of iPfam have structural evidence of interaction and hence the rate of false positives would be less. Thus, prediction of an interaction between proteins of host and pathogen genomes using different template databases can lead to more confidence in the prediction of a particular interaction.

The method proposed to detect interaction between host-guest interactions can be applied to any organism if we know the gross details of the cell types the guest resides in. For example, it is known that *Mycobacterium tuberculosis* infects the macrophages of human lymph nodes and resides in them to cause tuberculosis. The analysis of *M. tuberculosis* interactions with human proteins can be focused on human proteins known to be expressed in the macrophages and thus generate hypotheses about the pathogenesis of *M. tuberculosis* [O. Krishnadev, S. Bisht, N. Chitra & N. Srinivasan, manuscript in preparation].

Conclusion

A data integration method to detect the protein-protein interactions between two genomes one of which resides inside the other organism is proposed in this paper. The method has been applied to detect protein-protein interactions between the *P. falciparum* and human host. Homology-based methods were used to find a superset of all interactions which are physically and structurally compatible. These putative *in vitro* interactions are expected to include many biologically feasible interactions. In order to narrow down further to identify biologically feasible interactions, data sets from different sources have been taken into consideration and it was found that due to the integration of expression and sub-cellular localization data, many physically viable interactions are removed as they are unlikely to occur in the biological context. The remaining interactions were subjected to a manual inspection which resulted in the identification of many biologically meaningful and significant interactions potentially responsible

for the pathogenesis. Thus, the method proposed is shown to be effective in detecting biologically meaningful protein-protein interactions, although the coverage of possible interactions may not be exhaustive and will improve with growth in data size.

Acknowledgements

We thank Dr. Rekha and Mr. Sairam Swamy for providing a dataset of protein-protein interactions of known 3-D structure. O. K. is supported by Council of Scientific and Industrial Research, New Delhi. This research is supported partly by NMITLI project from CSIR, India and computational genomics project funded by the Department of Biotechnology, India.

References

- Aloy, P., Ceulemans, H., Stark, A. and Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. J. Mol. Biol. 332, 989-998.
- Altieri, D. C. and Edgington, T. S. (1988). The saturable high affinity association of factor X to ADP-stimulated monocytes defines a novel function of the Mac-1 receptor. J. Biol. Chem. 263, 7007-7015.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
- Bahl, A., Brunk, B., Crabtree, J., Fraunholz, M. J., Gajria, B., Grant, G. R., Ginsburg, H., Gupta, D., Kissinger, J. C., Labo, P., Li, L., Mailman, M. D., Milgram, A. J., Pearson, D. S., Roos, D. S., Schug, J., Stoeckert, C. J. Jr. and Whetzel, P. (2003). PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. Nucleic Acids Res. 31, 212-215.
- Ben Mamoun, C., Gluzman, I. Y., Hott, C., MacMillan, S. K., Amarakone, A. S., Anderson, D.

L., Carlton, J. M.-R., Dame, J. B., Chakrabarti, D., Martin, R. K., Brownstein, B. H. and Goldberg, D. E. (2001). Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite *Plasmodium falciparum* revealed by microarray analysis. Mol. Microbiol. 39, 26-36.

- Bozdech, Z., Zhu, J., Joachimiak, M. P., Cohen, F. E., Pulliam, B. and DeRisi, J. L. (2003). Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. Genome Biol. 4, R9.

- Davis, F. P., Barkan, D. T., Eswar, N., McKerrow, J. H. and Sali, A. (2007). Host-pathogen protein interactions predicted by comparative modeling. Protein Sci. 16, 2585-2596

- Deng, M., Lancto, C. A. and Abrahamsen, M. S. (2004). *Cryptosporidium parvum* regulation of human epithelial cell gene expression. Int. J. Parasitol. 34, 73-82.

- Dessens, J. T., Beetsma, A. L., Dimopoulos, G., Wengelnik, K., Crisanti, A., Kafatos. F. C. and Sinden, R. E. (1999). CTRP is essential for mosquito infection by malaria ookinetes. EMBO J. 18, 6221-6227.

- Dyer, M. D., Murali, T. M. and Sobral, B. W. (2007). Computational prediction of host-pathogen protein-protein interactions. Bioinformatics 23, i159-166.

- Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. 300, 1005-1016.

- Finn, R. D., Marshall, M. and Bateman, A. (2005). iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics 21, 410-412.

- Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. and Bateman, A. (2006). Pfam: clans, web tools and services. Nucleic Acids Res. 34, D247-251.

- 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., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R.E., Yates, J. R. and Carucci, D. J. (2002). A proteomic view of the *Plasmodium falciparum* life cycle. Nature 419, 520-526.

- ...sner, E. R. and North, S. C. (1999). An open graph visualization system and its applications to software engineering. Softw. Pract. Exper. 00S1, 1-5.

- Gardy, J. L., Laird, M. R., Chen, F., Rey, S., Walsh, C. J., Ester, M. and Brinkman, F. S. (2005). PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21, 617-623.

- Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., *et al.*. (2003). A protein interaction map of *Drosophila melanogaster*. Science 302, 1727-1736.

- Goel, V. K., Li, X., Chen, H., Liu, S.-C., Chishti, A. H. and Oh, S. S. (2003). Band 3 is a host receptor binding merozoite surface protein 1 during the *Plasmodium falciparum* invasion of erythrocytes. Proc. Natl. Acad. Sci. USA 100, 5164-5169.

- Gowri, V. S., Krishnadev, O., Swamy, C. S. and Srinivasan, N. (2006). MulPSSM: a database of multiple position-specific scoring matrices of protein domain families. Nucleic Acids Res. 34, D243-246.

- Han, K., Park, B., Kim, H., Hong, J. and Park, J. (2004). HPID: the Human Protein Interaction Database. Bioinformatics 20, 2466-2470.

- Hiller, N. L., Bhattacharjee, S., van Ooij, C., Liolios, K., Harrison, T., Lopez-Estraño, C. and Haldar, K. (2004). A host-targeting signal in virulence proteins reveals a secretome in malarial infection. Science 306, 1934-1937.

- Hirokawa, T., Boon-Chieng, S. and Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 14, 378-379.

- Hoffman, S. L., Subramanian, G. M., Collins, F. H. and Venter, J. C., (2002). *Plasmodium*, human and *Anopheles* genomics and malaria. Nature 415, 702-709.

- Huang, T.-W., Tien, A.-C., Huang, W.-S., Lee, Y.-C., Peng, C.-L., Tseng, H.-H., Kao, C.-Y. and Huang, C.-Y. F. (2004). POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. Bioinformatics 20, 3273-3276.

- Imanishi, T., *et al.* (2004). Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol. 2, e162.

- Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interaction. Proc. Natl. Acad. Sci. USA 98, 4569-4574.

- Jenner, R. G. and Young, R. A. (2005). Insights into host responses against pathogens from transcriptional profiling. Nat. Rev. Microbiol. 3, 281-294.

- Kappe, S. H., Kaiser, K. and Matuschewski, K. (2003). The *Plasmodium* sporozoite journey: a rite of passage. Trends Parasitol. 19, 135-143.

- Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J. Mol. Biol. 305, 567-580.

- LaCount, D. J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J. R., Schoenfeld, L. W., Ota, I., Sahasrabudhe, S., Kurschner, C., Fields, S. and Hughes, R. E. (2005). A protein interaction network of the malaria parasite *Plasmodium falciparum*. Nature 438, 103-107.

- 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. and Mann, M. (2002). Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. Nature 419, 537-542.

- Le Bonniec, S., Deregnaucourt, C., Redeker, V., Banerjee, R., Grellier, P., Goldberg, D. E. and Schrével, J. (1999). Plasmepsin II, an acidic hemoglobinase from the *Plasmodium falciparum* food vacuole, is active at neutral pH on the host erythrocyte membrane skeleton. J. Biol. Chem. 274, 14218-14223.

- Legrain, P., Wojcik, J. and Gauthier, J. M. (2001). Protein--protein interaction maps: a lead towards cellular functions. Trends Genet. 17, 346-352.

- Le Roch, K. G., Zhou, Y., Batalov, S., and Winzeler, E. A. (2002). Monitoring the chromosome 2 intraerythrocytic transcriptome of Plasmodium falciparum using oligonucleotide arrays. Am. J. Trop. Med. Hyg. 67, 233-243.

- Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., *et al.* (2004). A map of the interactome network of the metazoan *C. elegans*. Science 303, 540-543.

- Li, X., Chen, H., Oo, T. H., Dal, T. M., Bergman, L. W., Liu, S.-C., Chishti, A. H. and Oh, S. S. (2004). A co-ligand complex anchors *Plasmodium falciparum* merozoites to the erythrocyte invasion receptor band 3. J. Biol. Chem. 279, 5765-5771.

- Loregian, A. and Palù, G. (2005). Disruption of protein-protein interactions: towards new targets for chemotherapy. J. Cell. Physiol. 204, 750-762.

- Mangeat, P., Roy, C. and Martin, M. (1999). ERM proteins in cell adhesion and membrane dynamics. Trends Cell Biol. 9, 187-192.

- 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.

- Merrell, D. S. and Falkow, S. (2004). Frontal and stealth attack strategies in microbial pathogenesis. Nature 430, 250-256.

- Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.

- Nair, R. and Rost, B. (2002). Sequence conserved for subcellular localization. Protein Sci. 11, 2836-2847.

- Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., Florio, A. and Cesareni, G. (2005). HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. BMC Bioinformatics 6 Suppl 4, S21.

- Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schächter, V., Chemama, Y., Labigne, A. and Legrain, P. (2001). The protein-protein interaction map of *Helicobacter pylori*. Nature 409, 211-215.

- Randi, A. M. and Hogg, N. (1994). I domain of beta 2 integrin lymphocyte function-associated antigen-1 contains a binding site for ligand intercellular adhesion molecule-1. J. Biol. Chem. 269, 12395-12398.

- Rekha, N., Machado, S. M., Narayanan, C., Krupa, A. and Srinivasan, N. (2005). Interaction interfaces of protein domains are not topologically equivalent across families within superfamilies: Implications for metabolic and signaling pathways. Proteins 58, 339-353.

- Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A. and Chinnaiyan, A. M. (2005). Probabilistic model of the human protein-protein interaction network. Nat. Biotechnol. 23, 951-959.

- Ridley, R. G. (2002). Medical need, scientific opportunity and the drive for antimalarial drugs. Nature 415, 686-693.

- 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.

- Sargeant, T. J., Marti, M., Caler, E., Carlton, J. M., Simpson, K., Speed, T. P. and Cowman, A. F. (2006). Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. Genome Biol. 7, R12.

- Singh, S., Kennedy, M. C., Long, C. A., Saul, A. J., Miller, L. H. and Stowers, A. W. (2003). Biochemical and immunological characterization of bacterially expressed and refolded *Plasmodium falciparum* 42-kilodalton C-terminal merozoite surface protein 1. Infect. Immun. 71, 6766-6774.

- Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y. and Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. Proc. Natl. Acad. Sci. USA 104, 4337-4341.

- Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H. and Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell 122, 957-968.

- Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. and Hogenesch, J. B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA 101, 6062-6067.

- Sultan, A. A., Thathy, V., Frevert, U., Robson, K. J., Crisanti, A., Nussenzweig, V., Nussenzweig, R. S. and Ménard, R. (1997). TRAP is necessary for gliding motility and infectivity of plasmodium sporozoites. Cell 90, 511-522.

- Taraschi, T. F., O'Donnell, M., Martinez, S., Schneider, T., Trelka, D., Fowler, V. M., Tilley, L. and Moriyama, Y. (2003). Generation of an erythrocyte vesicle transport system by *Plasmodium falciparum* malaria parasites. Blood 102, 3420-3426.

- Tian, W. and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol. 333, 863-882.

- Ueda, T., Rieu, P., Brayer, J. and Arnaout, M. A. (1994). Identification of the complement iC3b binding site in the β2 integrin CR3 (CD11b/CD18). Proc. Natl. Acad. Sci. USA 91, 10680-10684.

- Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403, 623-627.

- Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Tatusova, T. A. and Wagner, L. (2003). Database resources of the National Center for Biotechnology. Nucleic Acids Res. 31, 28-33.

- Wickham, M. E., Culvenor, J. G. and Cowman, A. F. (2003). Selective inhibition of a two-step egress of malaria parasites from the host erythrocyte. J. Biol. Chem. 278, 37658-37663.

- Wojcik, J., Boneca, I. G. and Legrain, P. (2002). Prediction, assessment and validation of protein interaction maps in bacteria. J. Mol. Biol. 323, 763-770.

- Zhong, W. and Sternberg, P. W. (2006). Genome-wide prediction of *C. elegans* genetic

interactions. Science 311, 1481-1484.

- Zutshi, R., Brickner, M. and Chmielewski, J. (1998). Inhibiting the assembly of protein-protein interfaces. Curr. Opin. Chem. Biol. 2, 62-66.