Published in final edited form as:

Methods. 2015 July 15; 83: 63–70. doi:10.1016/j.ymeth.2015.04.002.

Heavy Path Mining of Protein-Protein Associations in the Malaria Parasite

Xinran Yu,

Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249, USA, Phone: (210) 458-6493

Turgay Korkmaz,

Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249, USA, Phone: (210) 458-7346

Timothy G. Lilburn,

Novozymes NA, Durham, NC 27709, USA, Phone: (919) 494-3913

Hong Cai,

Department of Biology, South Texas for Emerging Infectious Diseases, University of Texas at San Antonio, San Antonio, TX 78249, USA, Phone: (210)458-6492

Jianying Gu, and

Department of Biology, College of Staten Island, City University of New York, Staten Island, NY 10314, USA, Phone: (718) 982-4123

Yufeng Wang

Department of Biology, South Texas for Emerging Infectious Diseases, University of Texas at San Antonio, One UTSA Circle, San Antonio, Texas 78249, USA, Phone: (210)458-6492, Fax: (210)458-5658

Xinran Yu: xinran.yu@utsa.edu; Turgay Korkmaz: turgay.korkmaz@utsa.edu; Timothy G. Lilburn: tglilburn@gmail.com; Hong Cai: hong.cai@utsa.edu; Jianying Gu: jianying.gu@csi.cuny.edu; Yufeng Wang: yufeng.wang@utsa.edu

Abstract

Correspondence to: Yufeng Wang, yufeng.wang@utsa.edu.

Author's contributions

YW and XY conceived and designed the study. XY and TK designed and implemented algorithms. TGL, HC, JG, and YW performed bioinformatic data mining and genomic annotation. All authors drafted the manuscript. All authors read and approved the final manuscript.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences, National Institute of Allergy and Infectious Diseases, or the National Institutes of Health.

The URL of the source code for super-sequence mining method is <http://www.cs.utsa.edu/~korkmaz/research/heavy-path-mining/>.

Statement

An abridged version of this paper appears in the Proceeding of the IEEE International Conference on Bioinformatics and Biomedicine Belfast, UK: IEEE; 2014: 107–112.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Annotating and understanding the function of proteins and other elements in a genome can be difficult in the absence of a well-studied and evolutionarily close relative. The causative agent of malaria, one of the oldest and most deadly global infectious diseases, is a good example of this problem. The burden of malaria is huge and there is a pressing need for new, more effective antimalarial strategies. However, techniques such as homology-dependent annotation transfer are severely impaired in this parasite because there are no well-understood close relatives. To circumvent this approach we developed a network-based method that uses a heavy path network-mining algorithm. We uncovered the protein-protein associations that are implicated in important cellular processes including genome integrity, DNA repair, transcriptional regulation, invasion, and pathogenesis, thus demonstrating the utility of this method.

Keywords

Bioinformatics; drug target; heavy path mining; genomics; malaria; systems biology

Introduction

It is notoriously difficult to study the biology of parasites, none more so than the Apicomplexans. Their life cycle is complex, including multiple hosts and multiple locations within these hosts, so *in vitro* studies are challenging. This is of more than academic interest because the disease caused by the members of the Apicomplexan genus *Plasmodium*, malaria, is one of the most devastating infectious diseases. It is estimated to have caused 584,000 deaths in 2013 alone and the 198 million cases that year had profound effects on the economies of many tropical and sub-tropical countries. While drugs against malaria exist, the malaria parasites are able to quickly evolve resistance to multiple drug treatments and rapidly adapt to changes in the host environment. The search for new antimalarial targets received a boost from the development of cutting-edge -omics technologies, starting with the reconstruction of the genome sequences of *Plasmodium falciparum*, the causative agent of the most lethal form of malaria, and its sibling species [1–7]. The expectation that this would lead to rapid breakthroughs was not met, however, as the characteristics of the genome have proven to be almost as unusual as the parasite's life cycle. More than 60% of the open reading frames (ORFs) of *P. falciparum* were annotated as “hypothetical proteins” without functional classification, due to the remote evolutionary relationships between the malaria parasite and other organisms, a problem that is compounded by the extremely high percentage of A and T residues in the sequence. These characteristics have frustrated attempts to use traditional sequence-homology based annotation approaches and thus slowed the search for drug targets and vaccines.

Sequence-based methods attempt to assign a function to a protein via homology – similar sequences encode similar functions. But it should also be possible to assign functions to proteins by looking at the proteins it is associated with, whether in metabolic pathways or in mechanisms involved in cell entry, for example. This necessitates a more holistic view of the entity we are trying to understand. Systems biology, which promotes such holistic views, has emerged as a new paradigm in malaria research, focusing on the characterization of complex interactions and control mechanisms of parasite biology and parasite-host

interactions [8–11]. This paradigm has been enabled by the integration of various high throughput data: temporal- and spatial- specific expression profiles have been revealed by customized microarray, RNA-Seq, and proteomic assays [12–27]; genome-wide association analyses have begun to unveil the loci that are associated with variability in multi-drug resistance among clinical parasite strains [28, 29]. Taking advantage of the interactions and associations that the integration of these data has created, we previously developed a neighborhood subnetwork alignment approach to predict the proteins that are associated with transcriptional regulation and cell cycle regulation [30, 31]. In this paper, we present a heavy path mining algorithm [32] to identify potentially important protein-protein associations in the malaria parasite *P. falciparum*. This method can be extended to other organisms and provides an alternate route to understanding the function of proteins in the context of the cell when sequence-based methods fail.

Methods

1. Graph model

In this section we introduce the notation used in this paper and formally define the heavy protein chain problem. We assume that a protein association network is denoted by $G=(V, E)$, where $V = \{v_1, v_2, \dots, v_n\}$, which is the set of all the proteins in the network, and $E = \{e_1, e_2, \dots, e_m\}$, the set of all edges. Each edge e_i connects v_a and v_b if and only if there is an interaction between protein v_a and protein v_b . We downloaded the protein-protein association data for *P. falciparum* from the STRING database [33]. The weight of edge e_i is set to be the confidence score of the protein-protein association, ranging from 0 to 1, based on sequence similarity, pathway assignment and analysis [34], text mining, genome organization, and evolutionary relationship.

Let $p = \{v_1, v_2, \dots, v_k\}$ be a chain with length $(k-1)$. The heaviest protein chain problem can be abstracted as a heaviest path problem in a graph. Based on [35], we have the following definition:

Definition 1—Given a graph G , a length k , and a threshold δ , the **heavy path mining** problem is to find the set $P = \{p_1, p_2, \dots, p_r\}$, where each path $p_j = \{v_1, v_2, \dots, v_k\}$ satisfies the following condition:

$$Support(p_j) = \sum_{i=1}^{k-1} Support(v_i v_{i+1}) > \delta.$$

An example of a directed graph G is shown in Fig. 1, which can be represented by an adjacency matrix M_1 (Fig. 2).

The adjacency matrix denoted by M_1 can be directly computed from the given graph G . The number of $M_1[v_i][v_j]$ is the weight on the edge $v_i v_j$. The reason for using the notation M_1 is to indicate that $M_1[v_i][v_j]$ represents the support for 1-hop path (2-length sequence) from v_i to v_j . In the rest of the paper, we will continue to generalize this notation as M_k to store the supports for the k -hop heaviest path. Note that $M_k[v_i][v_j]$ will show the support for the k -hop

longest path starting with $v_i v_j$, not the path from v_i to v_j . The actual k -hop paths will be stored in a hash table.

In this study, we search for simple heavy paths (e.g., the path has no duplications of proteins). Therefore the heavy protein chains from the graph form a directed acyclic graph (DAG). In the next section, we will introduce the method of using dynamic programming to generate the heavy protein chains [36–39].

2. Algorithm

In this section we will explain how to use the Dynamic Programming method to compute the matrix M_k from M_1 and M_{k-1} . To generate the $(k+1)$ -length heavy paths, we attempted to compute all the k -hop heaviest paths whose total weights are greater than a given threshold. As mentioned above, the adjacency matrix $M_1[i][j]$ is the support for the 1-hop longest path from node i to node j . On the other hand, for $k > 1$, $M_k[i][j]$ is the support for the k -hop heaviest paths starting with link ij rather than being the weight of a path from i to j . At the same time, a hash table H_k that stores k -hop heaviest paths starting with ij is maintained. The keys in H_k are $[ij]$ for all the non-zero elements in M_k while the values are the k -hop heaviest paths starting with the corresponding ij . Here it is possible that multiple paths starting with ij have the same weights.

Considering the graph in Fig. 1, we illustrate the process of obtaining k -hop heaviest paths when k is 3, i.e., to compute M_3 and H_3 .

First, let us generate H_1 . This can be done by using the non-zero elements from the original one-step sequence matrix M_1 . Resulting keys and 1-hop path in H_1 are:

$$H_1 = \begin{bmatrix} \text{Key:} & \text{Value} \\ AB: & \langle AB \rangle \\ BC: & \langle BC \rangle \\ BE: & \langle BE \rangle \\ CD: & \langle CD \rangle \\ ED: & \langle ED \rangle \end{bmatrix}$$

The next step is to compute M_2 by using M_1 twice as follows. For each i and j , we first search the largest value in row j of M_1 and identify its column index as $\max C$. Then $M_2[i][j]$ is computed by taking the summation of $M_1[i][j]$ and $M_1[j][\max C]$. If $M_1[i][j]$ is 0, we keep it 0 in $M_2[i][j]$. Also, if the maximum in row j of M_1 is 0, $M_2[i][j]$ is set to 0. The resulting M_2 is shown in Fig. 3.

While M_2 is being computed, we can generate the corresponding hash table H_2 and ensure that the paths in the hash table will not have loops. After $M_2[i][j]$ is computed, we check if it is 0. If it is not 0, we copy the value of the key $[j\max C]$ in H_1 to $[ij]$'s value in H_1 after the first element. For example, the new value in H_2 for key $[BC]$ will be $\langle BCD \rangle$. This is obtained by copying $\langle BC \rangle$ to $\langle CD \rangle$ after C since $M_1[B][E]$ has the largest value in row C. The completed H_2 is:

$$H_2 = \begin{bmatrix} \text{Key:} & \text{Value} \\ AB: & \langle ABE \rangle \\ BC: & \langle BCD \rangle \\ BE: & \langle BED \rangle \\ EB: & \langle EBC \rangle \end{bmatrix}$$

This shows that 3-length super-sequence patterns are ABE, BCD, BED and EBC with the supports of 0.7, 0.6, 0.5, and 0.4, respectively.

Finally, we can compute M_3 by using M_1 and M_2 . For each i and j , the largest value in row j of M_2 is first searched and then its column index is identified as $maxC$. At last, $M_3[i][j]$ is computed by taking the summation of $M_1[i][j]$ and $M_2[j][maxC]$. The resulting M_3 is shown in Fig. 4.

In H_3 , $M_2[B][C]$ has the largest value in row B, so we copy the value of [BC]: $\langle BCD \rangle$ to the value of [AB]: $\langle ABE \rangle$, starting from the second element. Consequently, $\langle ABCD \rangle$ is obtained and saved in H_3 with the key value of [AB]. Using the same method, $\langle EBCD \rangle$ can be generated. Thus we have $\langle ABCD \rangle$ and $\langle EBCD \rangle$ in hash table H_3 :

$$H_3 = \begin{bmatrix} \text{Key:} & \text{Value} \\ AB: & \langle ABCD \rangle \\ EB: & \langle EBCD \rangle \end{bmatrix}$$

In summary, we computed ABE as the 2-hop heaviest path by combining 1-hop heaviest path AB and BE from M_1 . Since AB has the support of 0.4 and BE has the support of 0.3, the support of ABE is 0.7. Using the same method, we obtained ABCD as the 3-hop heaviest path with the support of 1.0. These procedures can be represented in a pseudo code form (Fig. 5). This algorithm was originally developed for directed graphs, and can be applied to the protein-protein association networks, which are generally viewed as undirected graphs, where M_1 is a diagonal symmetric matrix.

The complexity of our heuristic algorithm to compute the k -hop heaviest path is $O(n^4)$. This is because in computing the two-dimensional array M_k we need to go through each element in M_{k-1} , which takes $O(n^2)$. While seeking the maximum in row j of M_{k-1} is $O(n)$ and checking the hash tables and making sure there is no loss in each entry costs $O(n)$ in the worst case, the total running time is $O(n^4)$. In a real world dataset, which has a less dense set of relationships than a complete graph, the running time would be much less than $O(n^4)$ since there would be more and more 0s in M_k and the length of the hash table would be much smaller.

3. The summary of the workflow

Fig. 6 summarizes the input, output, and the workflow of the heavy path mining approach with its application to the *P. falciparum* protein association network. To apply the heavy path algorithm to the protein-protein association networks, we take each protein as a node in the graph and the confidence score between each two proteins as the weight of each edge. In

the next section we will show the experimental results of the heavy protein chains we found using the heavy path method.

Results and Discussion

The objective of this study is to use the proposed heavy path mining algorithm to search for the protein-protein associations with the strongest statistical support in the malaria parasite *Plasmodium falciparum*. The statistical support adopted in this study was based on the confidence scores in the STRING functional protein association database [33]. This database contains protein-protein association data for thousands of organisms and it is easy to extract the set of associations for a given organism. Each of the associations is supported by integrated data from experimental results and computational predictions and these associations form a network among the proteins of the organism, as many of them are associated with multiple proteins. While enormous amounts of data have been generated through high throughput experiments, including direct yeast 2-hybrid protein-protein physical interaction data [40, 41], little is known about how proteins interact with each other to coordinate complex processes such as development and invasion. To date, STRING remains one of the most important resources for integrated information about such interactions.

The proposed heavy path mining algorithm was able to identify ‘strongest links’ in *P. falciparum* networks, with varying weights (confidence scores) and lengths (number of nodes) (Table 1). The more stringent the threshold or/and the longer the protein chain allowed, the fewer protein chains are detected.

For example, the algorithm identified seven protein-protein association paths for length=5, and weight threshold = 3,800, including

1. MAL13P1.126, MAL13P1.118, MAL8P1.153, MAL7P1.87, PF08_0126
2. MAL13P1.126, MAL13P1.137, PF07_0016, MAL8P1.153, MAL13P1.118
3. MAL13P1.137, MAL13P1.118, MAL8P1.153, MAL7P1.87, PF08_0126
4. MAL13P1.215, MAL8P1.153, MAL13P1.137, MAL13P1.118, MAL13P1.126
5. MAL7P1.87, MAL8P1.153, MAL13P1.137, MAL13P1.118, MAL13P1.126
6. PF07_0016, MAL13P1.137, MAL8P1.153, MAL13P1.118, MAL13P1.126
7. PF08_0126, MAL7P1.87, MAL8P1.153, MAL13P1.118, MAL13P1.126

Clearly, some of the nodes in these seven chains are overlapping. Fig. 7 (generated using the Cytoscape program [42]) shows a combined network of the eight nodes that form the strongest links (See Table 2).

The most important biological question from the heavy chain mining is: what do these strongest associated proteins do in the parasite life cycle? Not surprisingly, four out of these eight proteins were annotated as conserved *Plasmodium* proteins of unknown function, which are not characterized or are under-characterized. However, these eight proteins may be associated with a large number of other proteins, revealed by gene co-expression, genome

context, KEGG pathway assignment, or wet lab validation. Table 3 lists representative proteins that are involved in the processes that control various aspects of parasite biology such as information processing, cell cycle regulation, cytoadherence, signaling, transcriptional regulation, transport, and virulence. Several of these proteins are localized to organelles that are potential drug targets including apicoplast and rhoptry.

In particular, four proteins predicted by our heavy path mining algorithms to have the strongest protein-protein associations may carry out important functions:

1. PF08_0126, also known as PF3D7_0803400, is a putative DNA repair protein RAD54. This family of proteins has been shown to play an integral role in double-strand break repair via homologous recombination. While its exact function is yet to be elucidated, it may be associated with 23 proteins (Fig. 8). It was shown to have direct physical interactions with two proteins, rhoptry-associated protein 1 (RAP1) (PF3D7_1410400) and a putative RNA binding protein (PF3D7_0515000), as revealed by yeast 2-hybrid (Y2H) experiments. In addition, its roles in DNA replication, DNA recombination, genome integrity, and cell cycle regulation are indicated by its interaction with a DNA polymerase epsilon, catalytic subunit a (PF3D7_0630300), a putative DNA topoisomerase III (PF3D7_1347100), a putative origin recognition complex subunit 2 (ORC2, PF3D7_0705300), two DNA helicases (PF3D7_1227100 and PF3D7_0918600), a putative ATP-dependent DNA helicase (PF3D7_1429900), a Rad51 homolog (PF3D7_1107400), a putative DNA repair exonuclease Mre11 (PF3D7_0107800), a cdc2-related protein kinase 4 (CRK4, PF3D7_0317200), and a NIMA related kinase 1 (NEK1, PF3D7_1228300). Its association with several hypothetical proteins suggests that it may be involved in yet uncharacterized functions.
2. MAL8P1.153, also known as PF3D7_0802100, is a putative transcription factor in the family of ApiAP2. ApiAP2 is becoming a promising antimalarial target class due to its demonstrated roles in parasite development [43]. It has a characteristic AP2 DNA-binding domain that is homologous to the plant *Apetala2* domain that controls floral development. Yeast 2-hybrid assays revealed that MAL8P1.153 has direct physical interaction with 32 proteins. It is associated with potentially important proteins including at least five serine-threonine protein kinases, several proteases such as hemolysin and subtilisins, a rhoptry-associated leucine zipper-like protein 1 (RALP1), a putative regulator of chromosome condensation, and rhoptry neck protein 4 (RON4) (Fig. 9), indicating that this putative ApiAP2 may be involved in cell signaling, parasite invasion, and parasite development.
3. MAL13P1.118, also known as PF3D7_1321500.2, is a putative 3',5'-cyclic nucleotide phosphodiesterase. This calmodulin-dependent enzyme may play an important role in signaling. Figure 10 shows the proteins associated with MAL13P1.118, many of which are hypothetical proteins. Proteomic analysis showed that it is expressed at the schizont stage as well as in the gametocytes.
4. MAL13P1.137 (PF3D7_1325400) is a hypothetical malaria antigen. Although its function remains almost completely unknown, it was predicted to have a large number of protein-protein associations (Fig. 11). Notably, it was shown by

microarray and proteomics to be highly expressed in the infectious schizont and merozoite stages in the red blood cycle.

Conclusions

Organisms that are not closely related to well-studied model organisms often prove difficult to annotate using standard methods. This is especially true for *Plasmodium* species. The barriers to annotation also extend to more recently developed systems biology approaches to understanding these organisms; it is a major challenge simply to find a starting point for network exploration. In this study, our goal was to develop a method for finding such starting points based on what is known about protein-protein associations in *Plasmodium falciparum*. To this end, we developed and applied a network mining approach to find and extend the heaviest, or most likely, protein-protein association paths in the malaria parasite. The algorithm led us to proteins that appear to be central players in the parasite life cycle. These proteins are associated with mechanisms ranging from DNA repair, DNA replication, transcription, and translation, to signal transduction, cell cycle progression, parasite invasion and virulence. All these processes present attractive targets for drug and vaccine development. On the algorithm side, a future direction is to extend it into a distributed and parallel version for potential large-scale omic applications.

Acknowledgments

We thank PlasmoDB for providing an all-in-one portal to the malaria genomic data. This work is supported by NIH grants GM100806 and AI080579 to YW. YW is also supported by NIH grant G12MD007591. We thank the Computational Biology Initiative at UTSA for providing computational support.

References

1. Carlton J. The *Plasmodium vivax* genome sequencing project. *Trends in parasitology*. 2003; 19 (5): 227–231. [PubMed: 12763429]
2. Carlton J, Silva J, Hall N. The genome of model malaria parasites, and comparative genomics. *Curr Issues Mol Biol*. 2005; 7(1):23–37. [PubMed: 15580778]
3. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, Crabtree J, Angiuoli SV, Merino EF, Amedeo P, et al. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature*. 2008; 455(7214):757–763. [PubMed: 18843361]
4. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, et al. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature*. 2002; 419(6906):512–519. [PubMed: 12368865]
5. Carlton JM, Escalante AA, Neafsey D, Volkman SK. Comparative evolutionary genomics of human malaria parasites. *Trends in parasitology*. 2008; 24(12):545–550. [PubMed: 18938107]
6. Pain A, Bohme U, Berry AE, Mungall K, Finn RD, Jackson AP, Mourier T, Mistry J, Pasini EM, Aslett MA, et al. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature*. 2008; 455(7214):799–803. [PubMed: 18843368]
7. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*. 2002; 419(6906):498–511. [PubMed: 12368864]
8. Isokpehi RD, Hide WA. Integrative analysis of intraerythrocytic differentially expressed transcripts yields novel insights into the biology of *Plasmodium falciparum*. *Malaria journal*. 2003; 2(1):38. [PubMed: 14617379]

9. Le Roch KG, Chung DW, Ponts N. Genomics and integrated systems biology in *Plasmodium falciparum*: a path to malaria control and eradication. *Parasite immunology*. 2012; 34(2–3):50–60. [PubMed: 21995286]
10. Winzeler EA. Applied systems biology and malaria. *Nat Rev Microbiol*. 2006; 4(2):145–151. [PubMed: 16362033]
11. Yeh I, Hanekamp T, Tsoka S, Karp PD, Altman RB. Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. *Genome research*. 2004; 14(5):917–924. [PubMed: 15078855]
12. Bozdech Z, Ginsburg H. Data mining of the transcriptome of *Plasmodium falciparum*: the pentose phosphate pathway and ancillary processes. *Malaria journal*. 2005; 4(1):17. [PubMed: 15774020]
13. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The Transcriptome of the Intraerythrocytic Developmental Cycle of *Plasmodium falciparum*. *PLoS Biol*. 2003; 1(1):E5. [PubMed: 12929205]
14. Bozdech Z, Mok S, Hu G, Imwong M, Jaidee A, Russell B, Ginsburg H, Nosten F, Day NP, White NJ, et al. The transcriptome of *Plasmodium vivax* reveals divergence and diversity of transcriptional regulation in malaria parasites. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105(42):16290–16295. [PubMed: 18852452]
15. Dharia NV, Bright AT, Westenberger SJ, Barnes SW, Batalov S, Kuhlen K, Borboa R, Federe GC, McClean CM, Vinetz JM, et al. Whole-genome sequencing and microarray analysis of ex vivo *Plasmodium vivax* reveal selective pressure on putative drug resistance genes. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107(46):20045–20050. [PubMed: 21037109]
16. Dharia NV, Plouffe D, Bopp SE, Gonzalez-Paez GE, Lucas C, Salas C, Soberon V, Bursulaya B, Kochel TJ, Bacon DJ, et al. Genome scanning of Amazonian *Plasmodium falciparum* shows subtelomeric instability and clindamycin-resistant parasites. *Genome research*. 2010; 20 (11): 1534–1544. [PubMed: 20829224]
17. Dharia NV, Sidhu AB, Cassera MB, Westenberger SJ, Bopp SE, Eastman RT, Plouffe D, Batalov S, Park DJ, Volkman SK, et al. Use of high-density tiling microarrays to identify mutations globally and elucidate mechanisms of drug resistance in *Plasmodium falciparum*. *Genome biology*. 2009; 10 (2):R21. [PubMed: 19216790]
18. Florens L, Liu X, Wang YF, Yang SG, Schwartz O, Peglar M, Carucci DJ, Yates JR, Wu YM. Proteomics approach reveals novel proteins on the surface of malaria-infected erythrocytes. *Molecular and biochemical parasitology*. 2004; 135(1):1–11. [PubMed: 15287581]
19. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, et al. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature*. 2002; 419(6906):520–526. [PubMed: 12368866]
20. Lasonder E, Ishihama Y, Andersen JS, Vermunt AMW, Pain A, Sauerwein RW, Eling WMC, Hall N, Waters AP, Stunnenberg HG, et al. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature*. 2002; 419(6906):537–542. [PubMed: 12368870]
21. Lasonder E, Janse CJ, van Gemert GJ, Mair GR, Vermunt AM, Douradinha BG, van Noort V, Huynen MA, Luty AJ, Kroeze H, et al. Proteomic profiling of *Plasmodium* sporozoite maturation identifies new proteins essential for parasite development and infectivity. *PLoS pathogens*. 2008; 4(10):e1000195. [PubMed: 18974882]
22. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder AA, Carucci DJ, et al. Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome research*. 2004; 14(11):2308–2318. [PubMed: 15520293]
23. Le Roch KG, Zhou Y, Batalov S, Winzeler EA. Monitoring the chromosome 2 intraerythrocytic transcriptome of *Plasmodium falciparum* using oligonucleotide arrays. *Am J Trop Med Hyg*. 2002; 67(3):233–243. [PubMed: 12408661]
24. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder AA, Batalov S, Carucci DJ, et al. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*. 2003; 301(5639):1503–1508. [PubMed: 12893887]

25. Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, Bohme U, Lemieux J, Barrell B, Pain A, Berriman M, et al. New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Molecular microbiology*. 2010; 76(1):12–24. [PubMed: 20141604]
26. Vignali M, Armour CD, Chen J, Morrison R, Castle JC, Biery MC, Bouzek H, Moon W, Babak T, Fried M, et al. NSR-seq transcriptional profiling enables identification of a gene signature of *Plasmodium falciparum* parasites infecting children. *J Clin Invest*. 2011; 121(3):1119–1129. [PubMed: 21317536]
27. Otto TD, Bohme U, Jackson AP, Hunt M, Franke-Fayard B, Hoeijmakers WA, Religa AA, Robertson L, Sanders M, Ogun SA, et al. A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC biology*. 2014; 12(1):86. [PubMed: 25359557]
28. Mu J, Myers RA, Jiang H, Liu S, Ricklefs S, Waisberg M, Chotivanich K, Wilairatana P, Krudsood S, White NJ, et al. *Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nature genetics*. 2010; 42(3):268–271. [PubMed: 20101240]
29. Su XZ, Wellems TE. Toward a high-resolution *Plasmodium falciparum* linkage map: Polymorphic markers from hundreds of simple sequence repeats. *Genomics*. 1996; 33(3):430–444. [PubMed: 8661002]
30. Cai H, Hong C, Gu J, Lilburn TG, Kuang R, Wang Y. Module-based subnetwork alignments reveal novel transcriptional regulators in malaria parasite *Plasmodium falciparum*. *BMC Syst Biol*. 2012; 6 (Suppl 3):S5. [PubMed: 23282319]
31. Cai H, Hong C, Lilburn TG, Rodriguez AL, Chen S, Gu J, Kuang R, Wang Y. A novel subnetwork alignment approach predicts new components of the cell cycle regulatory apparatus in *Plasmodium falciparum*. *BMC bioinformatics*. 2013; 14 (Suppl 12):S2. [PubMed: 24267797]
32. Yu, X.; Korkmaz, T.; Lilburn, TG.; Cai, H.; Gu, J.; Wang, Y. Heavy path mining reveals novel protein-protein associations in the malaria parasite *plasmodium falciparum*; The IEEE International Conference on Bioinformatics and Biomedicine; Belfast, UK: IEEE; 2014. p. 107–112.
33. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, et al. STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*. 2013; 41(Database issue):D808–815. [PubMed: 23203871]
34. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*. 2010; 38 (Database issue):D355–360. [PubMed: 19880382]
35. Yu, X.; Korkmaz, T. Super-sequence frequent pattern mining on sequential dataset; International Conference on Big Data; Santa Clara, CA, USA. 2013;
36. Bhattacharyya B, Dehne F. Using spine decompositions to efficiently solve the length-constrained heaviest path problem for trees. *Information Processing Letters*. 2008; 108(5):293–297.
37. Ioachim I, Gélinas S, Soumis F, Desrosiers J. A dynamic programming algorithm for the shortest path problem with time windows and linear node costs. *Networks*. 1998; 31(3):193–204.
38. Wu BY, Chao K-M, Tang CY. An efficient algorithm for the length-constrained heaviest path problem on a tree. *Information Processing Letters*. 1999; 69(2):63–67.
39. Zien, JY. Hierarchical string matching using multi-path dynamic programming. US Patent. 6,556,984. 2003.
40. LaCount DJ, Schoenfeld LW, Fields S. Selection of yeast strains with enhanced expression of *Plasmodium falciparum* proteins. *Molecular and biochemical parasitology*. 2009; 163(2):119–122. [PubMed: 19026694]
41. LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C, et al. A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature*. 2005; 438(7064):103–107. [PubMed: 16267556]
42. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T. A travel guide to Cytoscape plugins. *Nature methods*. 2012; 9(11):1069–1076. [PubMed: 23132118]

43. Painter HJ, Campbell TL, Llinas M. The Apicomplexan AP2 family: integral factors regulating Plasmodium development. *Molecular and biochemical parasitology*. 2011; 176(1):1–7. [PubMed: 21126543]

Highlights for

- We develop a heavy path network-mining algorithm for protein association networks.
- Using it, we predict potential central players in the malaria parasite life cycle.
- This method can be extended to other organisms for network mining and annotation.

Fig. 1.
An example of a directed graph.

$$M_1 = \begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} & \begin{pmatrix} 0 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0.3 \\ 0 & 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0.2 & 0 \end{pmatrix} \end{matrix}$$

Fig. 2.

The adjacency matrix for the graph in Fig. 1.

$$M_2 = \begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} & \begin{pmatrix} 0 & 0.7 & 0 & 0 & 0 \\ 0 & 0 & 0.6 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.4 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

Fig. 3.
 M_2 for the graph in Fig. 1.

$$M_3 = \begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} & \begin{pmatrix} 0 & 1.0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

Fig. 4.
 M_3 for the graph in Fig. 1.

Heuristic Algorithm Computing k-hop Heaviest Path**Input:** M_l , M_{k-l} , and H_k **Output:** M_k and H_k

```

for  $1 \leq i \leq n$  do
  for  $1 \leq j \leq n$  do
    if  $M_l[i][j] \neq 0$  then
      Find  $maxC$  such that
       $M_{k-l}[j]maxC \geq M_{k-l}[j][l]$  for  $1 \leq l \leq n$ 
      while there is a loop in the new path do
         $M_{k-l}[j]maxC = 0$ ,
        Find  $maxC$  such that
         $M_{k-l}[j]maxC > M_{k-l}[j][l]$  for  $1 \leq l \leq n$ 
      end while
      if  $M_{k-l}[j]maxC \neq 0$  then
         $M_k[i][j] = M_l[i][j] + M_{k-l}[j]maxC$  ,
        Update hash table  $H_k$ 
      else
         $M_k[i][j] = 0$ 
      end if
    end if
  end for
end for

```

Fig. 5.

The pseudo code of the heuristic algorithm proposed for searching heaviest paths.

Fig. 6. The workflow of heavy path mining of the *Plasmodium falciparum* protein association network.

Fig. 7.
The *P. falciparum* protein-protein association network with length 5 and weight >3,800.
Cytoscape was used for visualization.

Fig. 8. The proteins associated with a putative DNA repair protein RAD54 (PF08_0126, also known as PF3D7_0803400)

Network view was generated using Cytoscape. Node size is proportional to the degree of the connectivity of the node. Nodes are colored according to their functional classification. (J) Translation, ribosomal structure and biogenesis, (A) RNA processing and modification, (K) Transcription, (L) Replication, recombination and repair, (B) Chromatin structure and dynamics, (D) Cell cycle control, cell division, chromosome partitioning, (Y) Nuclear structure, (V) Defence mechanisms, (T) Signal transduction mechanisms, (M) Cell wall/membrane/envelope biogenesis, (N) Cell motility, (Z) Cytoskeleton, (W) Extracellular structures, (U) Intracellular trafficking, secretion, and vesicular transport, (O) Posttranslational modification, protein turnover, chaperones, (C) Energy production and conversion, (G) Carbohydrate transport and metabolism, (E) Amino acid transport and metabolism, (F) Nucleotide transport and metabolism, (H) Coenzyme transport and metabolism, (I) Lipid transport and metabolism, (P) Inorganic ion transport and metabolism, (Q) Secondary metabolites biosynthesis, transport and catabolism, (R) General function prediction only, and (S) Function unknown.

Fig. 9. The proteins associated with a putative ApiAP2 transcriptional regulator MAL8P1.153, also known as PF3D7_0802100.

Fig. 10.

The proteins associated with a putative 3',5'-cyclic nucleotide phosphodiesterase MAL13P1.118, also known as PF3D7_1321500.2.

Fig. 11.
The proteins associated with a conserved *Plasmodium* protein of unknown function MAL13P1.137, also known as PF3D7_1325400.

Table 1

The numbers of chains (N) with specific lengths (Len) found given threshold of weights (W)

Len = 5		Len = 6		Len = 7		Len = 8		Len = 9		Len = 10		Len = 11		Len = 12		Len = 13		Len = 14		Len = 15	
W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N	W	N
3500	236	4400	276	5300	336	6200	334	7100	238	7900	310	8800	164	9500	264	10400	85	11100	60	11900	13
3600	107	4500	158	5400	167	6300	148	7200	114	8000	161	8900	70	9600	137	10500	39	11200	27	12000	8
3700	41	4600	62	5500	71	6400	59	7300	33	8100	61	9000	16	9700	56	10600	17	11300	12	12100	5
3800	7	4700	12	5600	15	6500	11	7400	5	8200	9	9100	3	9800	18	10700	3	11400	2	12200	2
3900	0	4800	0	5700	1	6600	1	7500	0	8300	1	9200	0	9900	0	10800	0	11500	0	12300	1

Table 2

Protein-protein associations with length 5 with combined weights > 3,800

Protein 1	Annotation	Protein 2	Annotation	Confidence Score
MAL13P1.137	conserved Plasmodium protein, unknown function	MAL8P1.153	transcription factor with AP2 domain(s) (ApiAP2)	960
MAL13P1.118	cAMP-specific 3',5'-cyclic phosphodiesterase 4D, putative	MAL8P1.153	transcription factor with AP2 domain(s) (ApiAP2)	960
PF07_0016	conserved Plasmodium membrane protein, unknown function	MAL8P1.153	transcription factor with AP2 domain(s) (ApiAP2)	938
MAL13P1.215	conserved Plasmodium protein, unknown function	MAL8P1.153	transcription factor with AP2 domain(s) (ApiAP2)	948
PF07_0016	conserved Plasmodium membrane protein, unknown function	MAL13P1.137	conserved Plasmodium protein, unknown function	960
MAL7P1.87	conserved Plasmodium protein, unknown function	MAL8P1.153	transcription factor with AP2 domain(s) (ApiAP2)	960
MAL13P1.118	cAMP-specific 3',5'-cyclic phosphodiesterase4D, putative	MAL13P1.126	palmitoyltransferase, putative (DHHC5)	960
MAL13P1.118	cAMP-specific 3',5'-cyclic phosphodiesterase 4D, putative	MAL13P1.137	conserved Plasmodium protein, unknown function	960
MAL7P1.87	conserved Plasmodium protein, unknown function	PF08_0126	DNA repair protein rad54, putative	960
MAL13P1.137	conserved Plasmodium protein, unknown function	MAL13P1.126	palmitoyltransferase, putative (DHHC5)	948

Table 3

Representative proteins that are associated with the eight proteins with heaviest paths.

Protein accession number	Annotation
PF3D7_0728600	zinc finger, C3HC4 type, putative
PF3D7_0209600	transporter, putative
PF3D7_0220800	cytoadherence linked asexual protein 2 (CLAG2)
PF3D7_0501600	rhoptry-associated protein 2 (RAP2)
PF3D7_0501500	rhoptry-associated protein 3 (RAP3)
PF3D7_0506900	rhomboid protease ROM4 (ROM4)
PF3D7_0507200	subtilisin-like protease 3 (SUB3)
PF3D7_0604100	transcription factor with AP2 domain(s), SPE2-interacting protein (SIP2)
PF3D7_0530900	formin 1
PF3D7_0923500	cyclin-dependent kinase, putative
PF3D7_1238600	sphingomyelin phosphodiesterase, putative
PF3D7_0932200	profilin, putative (PFN)
PF3D7_0208000	serine repeat antigen 1 (SERA1)
PF3D7_0207700	serine repeat antigen 4 (SERA4)
PF3D7_0805700	serine/threonine protein kinase, FIKK family (FIKK8)
PF3D7_0803300	mitogen-activated protein kinase organizer 1, putative (PfMORG1)
PF3D7_1146600	CCAAT-box DNA binding protein subunit B (NFYB)
PF3D7_0705300	origin recognition complex subunit 2, putative (ORC2)
PF3D7_0107800	DNA repair exonuclease Mre11, putative
PF3D7_0816800	meiotic recombination protein DMC1, putative (DMC1)
PF3D7_0918600	DNA helicase, putative