Methods 83 (2015) 63–70

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Heavy path mining of protein–protein associations in the malaria parasite ☆

Xinran Yu [a], Turgay Korkmaz [a], Timothy G. Lilburn [b], Hong Cai [c], Jianying Gu [d], Yufeng Wang [c,*]

[a] Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249, USA
[b] Novozymes NA, Durham, NC 27709, USA
[c] Department of Biology, South Texas for Emerging Infectious Diseases, University of Texas at San Antonio, San Antonio, TX 78249, USA
[d] Department of Biology, College of Staten Island, City University of New York, Staten Island, NY 10314, USA

ARTICLE INFO

Article history:
Received 31 January 2015
Received in revised form 31 March 2015
Accepted 1 April 2015
Available online 7 April 2015

Keywords:
Bioinformatics
Drug target
Heavy path mining
Genomics
Malaria
Systems biology

ABSTRACT

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.

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

1. 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 subtropical 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,

☆ 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.

* Corresponding author. Fax: +1 (210)458 5658.

E-mail addresses: xinran.yu@utsa.edu (X. Yu), turgay.korkmaz@utsa.edu (T. Korkmaz), tglilburn@gmail.com (T.G. Lilburn), hong.cai@utsa.edu (H. Cai), jianying.gu@csi.cuny.edu (J. Gu), yufeng.wang@utsa.edu (Y. Wang).

http://dx.doi.org/10.1016/j.ymeth.2015.04.002
1046-2023/© 2015 Elsevier Inc. All rights reserved.

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.

2. Methods

2.1. 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 $V = \{v_1, v_2, ...v\}$, which is the set of all the proteins in the network, and $E = \{e_1, e_2, ..., e\}$, 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 $p = \{v_1, v_2, ..., v\}$ 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 $P = \{p_1, p_2, ...p\}$, where each path $p_j = \{v_1, v_2, ..., v\}$ satisfies the following condition:

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

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

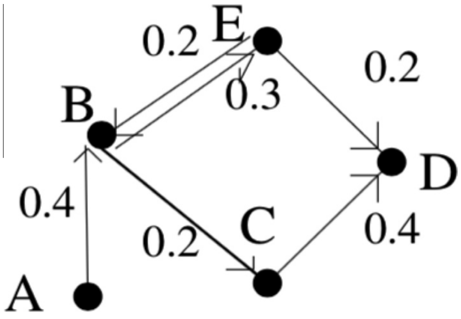

Fig. 1. An example of a directed graph.

$$M_1 = \begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 0 & 0.4 & 0 & 0 & 0 \\ B & 0 & 0 & 0.2 & 0 & 0.3 \\ C & 0 & 0 & 0 & 0.4 & 0 \\ D & 0 & 0 & 0 & 0 & 0 \\ E & 0 & 0.2 & 0 & 0.2 & 0 \end{array}$$

Fig. 2. The adjacency matrix for the graph in Fig. 1.

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].

2.2. 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:

$$M_2 = \begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 0 & 0.7 & 0 & 0 & 0 \\ B & 0 & 0 & 0.6 & 0 & 0.5 \\ C & 0 & 0 & 0 & 0 & 0 \\ D & 0 & 0 & 0 & 0 & 0 \\ E & 0 & 0.4 & 0 & 0 & 0 \end{array}$$

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