Clustering PPI Networks of Mixed Host-Pathogen Data Using Biased Repeated Random Walks

[1]Andrew Peterman, [2,3]Melanie J. Bennett, [2,3]Alan Frankel and Rahul Singh[1,4*]

[1]Department of Computer Science,
San Francisco State University, San Francisco, CA 94132
[2]The HARC Center, [3]Department of Biochemistry and Biophysics,
[4]Center for Discovery and Innovation in Parasitic Diseases,
University of California, San Francisco, San Francisco, CA 94158

Abstract. Clustering protein-protein interaction (PPI) network data yields groups of proteins that are biochemically involved. Most existing clustering methods treat all the proteins in a PPI network equally. However, analyzing host-pathogen networks requires identification of clusters that represent the interactions between the set of pathogen proteins and the set of host proteins. For studying HIV-human protein-protein interactions, we thus need to identify clusters with the specific composition of at least one virus protein per cluster. Towards this goal, we describe a novel clustering method that focuses on the key virus proteins in a host-pathogen PPI network and utilizes the notion of random walks biased towards specific connectivity configurations. The proposed method finds host-pathogen protein clusters with high accuracy and improves upon the results obtained with other methods at the state-of-the-art.

1 Introduction

Amongst PPI networks, host-pathogen PPI networks (HPPIN) represent a subclass that involves proteins from both a host, such as a human, and a pathogen genome such as HIV. HPPIN represent the mechanistic relationships that underlie biological interactions between organisms. Thus, such networks can be used to study phenomena such as the spread of viruses in a host or progression of a disease.

Many clustering algorithms have been developed and tested on PPI networks in the last decade. In most cases, the goal has generally been to separate and identify groups of proteins working together in normal cellular processes. In doing so, most network analysis methods treat nodes (proteins) homogeneously. Analysis of HPPIN focusing on the interaction of protein across species, however, requires treating the nodes (and consequent connectivity patterns) differentially since we are interested in understanding the inter-species protein-protein interactions. Algorithms designed specifically for the analysis of HPPINs can be especially valuable in identifying dependency factors - host proteins that are essential for the pathogen to spread but whose disruption may not significantly harm the host. Thus the problem constraints differ from those encountered in bipartite clustering.

In this paper we focus on analysis of HIV-Human HPPIN. Our HIV-human dataset consists of HIV-human PPIs from both HARC Proteomics [1] and VirusMINT [2] as

[*] Corresponding author; E-mail: rahul@sfsu.edu. This research was funded in part by the NSF grant IIS0644418 (CAREER)

well as human-human PPIs from HIPPIE[3]. The data includes over 13,000 unique proteins and more than 40,000 total interactions. The small number of HIV proteins (18), implies that an analysis technique would have to deal with highly imbalanced cluster distribution with most possible clusters comprising entirely of human proteins and consequently uninteresting given the problem formulation.

2 Method

Given the problem context, recently proposed techniques of interest include network flow based methods and random walk based methods. Conceptually, flow based methods work by simulating fluid flows in a network. Our research is based on the use of random walks. In this approach, an imaginary walker traverses a network by starting at a node and randomly taking steps to the neighboring nodes. Over time, the probability of ending a random walk at any particular node in the network converges. Furthermore, the traversal pattern highlights regions of high connectivity and can be used to identify clusters in a network. For a HPPIN such as ours with only 18 pathogen proteins, the majority of clusters returned through a direct application of the random walk strategy would not contain a pathogen protein and consequently would be of little help in understanding host-pathogen protein-protein interactions.

In clustering HPPIN, we would like an algorithm to focus on and identify only those clusters that involve at least one pathogen protein. In finding such clusters, if necessary, the method may need to prioritize weaker host-pathogen interactions over stronger interactions between proteins of the host. Being more inclusive of pathogen interactions, such an approach may identify interaction pathways that would be missed by traditional random walk-based methods.

The key steps of the proposed method which we call Biased Repeated Random Walk (BRRW) are:

- *Random walk-based network traversal:* First we create all the random walk stationary vectors x_i for all network nodes through a traversal process called random walk with restarts. A random walks with restarts, is a form of random walk that teleports the walker back to the starting node with some probability α.

  $$\boldsymbol{x_i} = \alpha \boldsymbol{s_i} + (1-\alpha)\boldsymbol{P^T} \boldsymbol{x_i} \qquad (1)$$

  The first term of equation 1 represents the starting node probability and the second term the flow of the walker away from the node according to the network topology. The stationary vectors x_i with resulting element probabilities (scores) $x_i[p]$ represent the converged probability values for a random walk starting from a network node n, where P is the transition matrix, α is the restart probability, and s_i is the starting node vector. In the following, we refer to scores of virus nodes as $x_i[p_v]$ and host nodes $x_i[p_h]$.
- *Biased cluster expansion*: For each node n the highest scoring $\boldsymbol{x}_i[p_v]$ is added to the cluster if it is a neighbor of n. Otherwise the highest host $x_i[p_h]$ is added to cluster C. If the first element n of C is a virus protein, then only host proteins will be added to C. Each x_i of the cluster is then averaged by summing each node's vector in the cluster and dividing by the cluster size to find x_c. The scores $x_c[p]$ represent the probabilities of reaching the node p from C. After the first virus

protein has been added, clustering continues to add the next highest $x_c[p]$ not already part of *C*.

- *Cluster expansion termination*: Clustering is complete when either the score $x_c[p]$ of the node being added to *C* is below a threshold value, or when $|C|$ reaches a predetermined size. It is possible that no virus protein is added to *C* before it is completed and in this case the resultant cluster is discarded.

3 Experimental Results

BRRW was run on our HPPIN network and compared against two other random walk methods [3, 4] as well as the SinkSource+ algorithm [6]. The resulting clusters of the random walk-based methods were compared for (1) the number of virus-containing clusters, and (2) number of distinct clusters per virus protein. These clusters were also compared with expert determined groupings to determine cluster quality. The total numbers of clusters found with the random walk-based methods were as follows: 704 using RW [4] (basic random walk from a node), 25 using RRW [5] (repeated random walk from a cluster of nodes), and 1396 using BRRW. The numbers of clusters per protein found using these methods is shown in Figure 1. In most cases BRRW returned many more clusters per virus protein than the other methods and exposed thereby a larger number of different interactions pathways. It may also be noted that while the number of clusters per protein obtained using RW and BRRW methods were similar, the cluster found using the RRW method were far fewer.

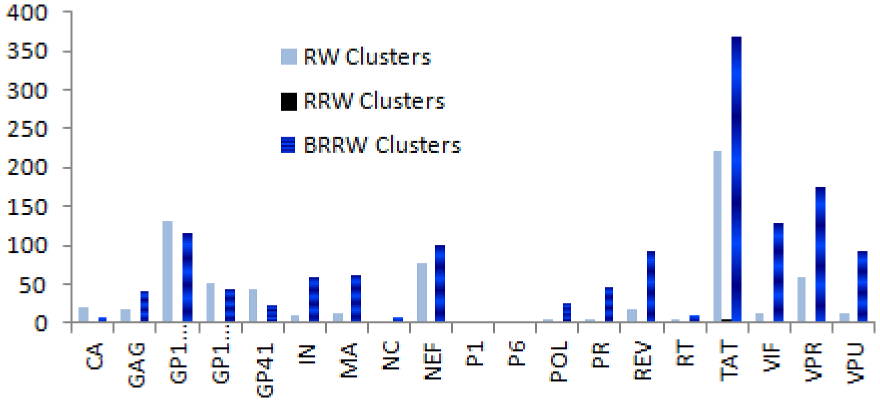

Figure 1. Number of distinct clusters per HIV protein obtained using three random walk-based methods. In addition to more clusters containing HIV proteins, the distinct number of proteins contained in the clusters is larger, giving more potential drug targets.

In the second experiment (see Table 1), three known clusters associated with HIV proteins from previous experiments[1] were used as gold standards to analyze the clustering results of the proposed method and compare it with those obtained with RW, RRW, and SinkSource+. These clusters involved interactions with the following three virus proteins respectively: VIF, VPR, TAT [1]. For each method, the most accurate cluster was determined by comparing all clusters to the gold standards using the Jacaard index. It may be noted, that the HIV protein in the gold standard was not always present in the best scoring cluster. As can be seen from the table, clusters obtained with the proposed method most closely matched the gold standard clusters in terms of composition. While limited in scale, this study illustrates the promise of the proposed method.

Method	Cluster 1	Score	Cluster 2	Score	Cluster 3	Score
Gold Standard	VIF, CUL5, TCEB1, TCEB2, RNF7,CBFB	-	VPR,GEMIN2, SMN1, DDX20, GEMIN6, GEMIN4,GEMIN7	-	TAT, CDK9, CCNT1, AF9, AFF1, AFF4, ELL, ENL, PAF1, LARP7, MEPCE	-
RW	VIF, CUL5, RNF7, TCEB1, TCEB2, RBX1, COPS6, NEDD8, NOS2, ASB2, VHL, DCUN1D1, CAND1, ASB4, ASB6	0.31	DDX20, SMN1, GEMIN4, GEMIN2, MYC, SNRPD2, SNRPE, GEMIN5, SNRPG, SNRPD1, SNRPB, EIF2C2, SNRPD3, SNRPF, SUMO2	0.22	TAT, CCNT1, CDK9, AFF1, HEXIM1, BRD4, AFF4, MYC, MLLT3, GRN, SNW1, MDFIC, MEPCE, AHR, PML	0.30
RRW	VIF, APOBEC3F, APOBEC3G	0.13	DDX20, GEMIN2, GEMIN5, SMN1, SNRPD1, SNRPD2, SNRPE, SNRPF, SNRPG	0.23	TAT, TAF1, TAF10, TAF13, TAF2, TAF4, TAF5, TAF6, TAF8, TAF9, TBP	0.05
SinkSource+	VIF, TCEB1	0.33	PHAX, GEMIN4, GEMIN5, GEMIN6,TGS1,GEMIN7, SNUPN, WDR77, DDX20, PRMT5	0.31	CDK9	0.09
BRRW	**VIF, CBFB, CUL5, TCEB1, TCEB2, RBX1, CUL2, RNF7**	**0.75**	**VPR, GEMIN2, SMN1, DDX20, GEMIN6, GEMIN4,GEMIN7**	**1.00**	**TAT, CDK9, CCNT1, SNW1, MYC, LARP7, AFF1, AFF4, ENL, HNRNPA1, MEPCE**	**0.57**

Table 1. Comparison of the most accurate clusters (scored by Jaacard index) returned by each method to biologically verified clusters. BRRW most closely matched the gold standards in terms of accuracy.

References

1. Jäger, S., Cimermancic, P., Gulbahce, N., *et al* (2012). Global Landscape of HIV-Human Protein Complexes. *Nature*, 481 (7381), 365-70.
2. Chatr-aryamontri A, Ceol A, Peluso D, Nardozza A, Panni S, et al. (2009) VirusMINT: a viral protein interaction database. *Nucleic Acids Research* 37: D669–673.
3. Schaefer MH, Fontaine J-F, Vinayagam A, Porras P, Wanker EE, et al. (2012) HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores. *PLoS ONE* 7(2): e31826.
4. Can T, Çamoğlu O, Singh AK: Analysis of protein-protein interaction networks using random walks. *Proceedings of the 5th International Workshop on Bioinformatics*, *pp*. 61 - 68, 2005.
5. Macropol K, Can T, Çamoğlu O, Singh AK: RRW: Repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics* 2009: 10:283
6. Murali TM, Dyer MD, Badger D, Tyler BM, Katze MG (2011) Network-Based Prediction and Analysis of HIV Dependency Factors. *PLoS Computational Biology* 7(9): e1002164.