Mining PPI Networks for Identifying Potential Pathways of Hepatitis-C Infection Leading to Various Diseases

Anirban Mukhopadhyay, *Senior Member, IEEE*, Ujjwal Maulik, *Senior Member, IEEE*.

***Abstract*—Protein-protein interaction network-based study of viral pathogenesis has been gaining popularity among computational biologists in recent days. In the present study we attempt to investigate the possible pathways of hepatitis-C virus (HCV) infection by integrating the HCV-human interaction network, human protein interactome and human genetic disease association network. We have proposed quasi-biclique and quasi-clique mining algorithms to integrate these three networks to identify infection gateway host proteins. Moreover, integrated study of the three networks reveals potential pathways of infection by the HCV that lead to various diseases including cancers. The gateway proteins are found to be biologically coherent and have high degrees in human interactome compared to the other virus-targeted proteins. The analyses also provide possible targets for more effective anti-hepatitis-C therapeutic involvement.**

***Index Terms*—Protein-Protein interaction network, hepatitis-C, quasi-clique, quasi-biclique, gene-disease association network.**

I. Introduction

Hepatitis-C virus (HCV) causes the infectious disease Hepatitis-C which primarily affects the liver. It is important to identify the potential target human proteins that lead to different diseases caused by hepatitis-C virus infection. Analyzing the regulation between viral and host proteins in different organisms helps to uncover the underlying mechanism of various viral diseases. Protein-protein interaction (PPI) information provides a local as well as a global view of the interaction modules of proteins participating in similar biological activities. Such interaction information can be obtained via biological experimentations or can be predicted using computational approaches [1].

One of the main goals in research of PPI is to predict possible viral-host interactions. This interaction information can be utilized to identify and prioritize the important viral-host interactions. This is specifically aimed at assisting drug developers targeting protein interactions for the development of specially designed small molecules to inhibit potential HCV-Human PPIs. Targeting protein-protein interactions has relatively recently been established to be a promising alternative to the conventional approach to drug design [2].

Although there have been many studies on determining and analyzing PPIs in a single organism, not much work can be found on computational analysis of viral-host interactions. In very recent times, some computational analysis of viral-host interactions, specially in HIV-1-human PPIs [3]–[7] have been done. Although some recent studies have analyzed the viral-host interactions for some individual HCV proteins [8], [9], no global system-wide study based on the HCV-human interaction network is available in literature. Motivated by this, in the present work, the PPI records between HCV proteins and human (*Homo sapiens*) proteins reported in a recently published dataset [10] are collected. This interaction information, all together, can be visualized as a bipartite graph, where two sets of nodes denote HCV proteins and human proteins, respectively, and the edges denote the interactions. In this work, the bipartite network is mined to identify the strong interacting modules, which are effectively quasi-bicliques. We further extend the study by clustering the human protein-protein interaction network (PPIN) to identify the possible quasi-cliques that overlap with the quasi-bicliques identified in the previous step. The human proteins participating in these quasi-cliques are considered as gateways of infection and are investigated for their functional characteristics. Subsequently, the bipartite network representing the association of human proteins with various disease types is mined to find possible quasi-bicliques that overlap with the gateway proteins discovered in the previous stage. Thus we explore three networks, namely, HCV-human PPIN, human PPIN, and human proteins-disease association network to discover the potential infection pathways of HCV that lead to various diseases including cancers. The analyses done in this study may provide possible targets for more effective anti-hepatitis-C drugs.

II. Proposed Study

In the present study, three different networks are mined. First one is the HCV-human protein interaction bipartite network. The second network is human PPIN, which is modeled as a graph. The third network represent the associations between human proteins and disease. Hence this disease association network is also modeled as a bipartite graph with two sets of nodes representing human proteins and diseases, respectively.

The proposed study consists of three stages. First we mine strong γ-quasi-bicliques from the first bipartite graph representing the interactions between viral and human proteins. A γ-quasi-biclique is defined as dense bipartite subgraph with density at least γ ($0 \leq \gamma \leq 1$). The obtained quasi-bicliques

A. Mukhopadhyay is with the Department of Computer Sc. & Engg., University of Kalyani, Kalyani-741235, India. E-mail: anirban@klyuniv.ac.in

U. Maulik is with the Department of Computer Sc. & Engg., Jadavpur University, Kolkata-700032, India. E-mail: umaulik@cse.jdvu.ac.in

The codes, datasets and other related materials can be obtained from the following website: www.anirbanm.in/hcv

are strong interaction modules consisting of the HCV and human proteins. Thereafter, in the second stage we cluster the human PPIN to identify possible strong γ-quasi-cliques that overlap with the quasi-bicliques identified in the previous stage. A γ-quasi-clique is defined as a subgraph having density at least γ ($0 \leq \gamma \leq 1$). The human proteins participating in these quasi-cliques are considered as gateways of infection and are further investigated for their functional characteristics. Subsequently, the bipartite network representing the association of human proteins with various disease types is mined to find possible strong γ-quasi-bicliques overlapping with the gateway proteins discovered in the previous stage. Hence we explore three networks, namely, HCV-human PPIN, human PPIN, and human proteins-disease association network to discover the potential pathways of HCV infection that lead to various diseases including cancers.

Here we have proposed an algorithm which can mine both γ-quasi-cliques and γ-quasi-bicliques from graphs and bipartite graphs, respectively. The algorithm is basically a quasi-clique mining algorithm which can also be used to mine quasi-bicliques with little modification. First we describe the algorithm for mining quasi-cliques. Thereafter, how this algorithm is modified to mine quasi-bicliques is described.

A. Mining γ-quasi-cliques

The proposed algorithm for mining γ-quasi-cliques is based on hierarchical clustering method [11]. Given an input graph $G = (V, E)$, first the shortest path distances (number of edges) between all vertex-pairs are computed. Thereafter a dendrogram is built using agglomerative average linkage method as follows: First for each vertex, a cluster is formed. Subsequently two nearest vertices are combined to form a new cluster. This continues until there remains only one cluster containing all the vertices. The distance between any two clusters is computed as the average distance among all the vertices in the two clusters. The tree representing the hierarchical relationships among the clusters is called a dendrogram.

After that, we start scanning from the top of the dendrogram to the bottom, one level at a time. Each time a cluster is divided into two, we examine the two clusters whether they are γ-quasi-cliques given a γ value. If any cluster satisfies this criterion, we do not further divide that cluster, i.e., the subtree rooted by this cluster is no more explored and this cluster is returned as one γ-quasi-clique. The clusters that are not γ-quasi-cliques are recursively divided as per the dendrogram until they provide some γ-quasi-clique, or reaches the threshold of quasi-clique size (minimum number of vertices to be present in the quasi-clique). Hence, the algorithm returns a set of maximal γ-quasi-cliques, i.e., the γ-quasi-cliques which are not completely included in another γ-quasi-clique.

B. Mining γ-quasi-bicliques

The algorithm for mining γ-quasi-bicliques is exactly same as mining γ-quasi-cliques, the only modification is done in the distance matrix. In this case also, we compute the shortest path between the nodes in the input bipartite graph $G = (V_1, V_2, E)$. Note that here the distance between two vertices $u \in V_1$ and $v \in V_2$ can be any odd value ≥ 1, since u and v may not be directly connected, but there may be a path between this two that contains a number of vertices from V_1 and v_2 in alternative positions. Any two vertices $u_1, u_2 \in V_1$ are never connected directly in a bipartite graph, however they may be connected through a set of vertices from V_2 and V_1 in an alternative fashion, and thus the distance between any two vertices in V_1 is always an even value ≥ 2. Similar is the case for any two vertices in set V_2.

The number of HCV proteins (set V_1) is much lesser than that of human proteins (set V_2). To increase the participation of HCV proteins in the γ-quasi-bicliques, we have modified the distance function between two viral proteins as follows: The distance between any two viral proteins that are connected by a series of alternative human and viral proteins, i.e., which belong to the same connected component in the bipartite graph, is set to 1. Thus the viral proteins belonging to the same connected component come closer to each other virtually and the number of viral proteins in the γ-quasi-bicliques increases. The same approach is adopted while finding the quasi-bicliques in the human protein-disease association network to increase the participation of the human proteins.

III. Databases and Preprocessing

As stated before, we deal with three networks, namely, HCV-human PPI network, human PPI network and human protein-disease association network. Here, the collection and preprocessing of the datasets have been described below.

A. HCV-Human Protein Interaction Database

The protein interaction information between the HCV proteins and human proteins have been collected from a recently developed HCV-human protein interaction database called HCVpro [10] publicly available at http://cbrc.kaust.edu.sa/hcvpro/. The database contains the interactions among 11 HCV proteins (CORE, E1, E2, F, NS2, NS3, NS4A, NS4B, NS5A, NS5B, p7) and 455 human proteins. The total number of interactions is 549. After removing the redundant interactions, the number of unique interactions reduces to 524. These 524 interactions among 11 HCV proteins and 455 human proteins are used for preparing the bipartite network between viral and host proteins and the maximal γ-quasi-bicliques are mined from this bipartite network as described in the previous section.

B. Human Protein Interaction Database

The primary objective of mining human protein interaction database is to find γ-quasi-cliques that overlap with the γ-quasi-bicliques identified in the previous stage of the study. Hence to avoid huge computational complexity in mining quasi-cliques from the complete human protein interaction database, we concentrate only on the part of the human PPI that contains the human proteins present in the identified γ-quasi-bicliques in the previous stage. For this, the function protein association network STRING (http://string-db.org/) has been utilized. For each quasi-biclique identified in the previous

stage, the participating human proteins are given as input to STRING and STRING generates an interactome containing these human proteins and other additional human proteins. We consider the predictions based on co-expression, experiments and databases only. We consider only the interactions with confidence of at least 0.8 (in a confidence scale between 0 and 1). This ensures that we consider only those PPIs that have reasonable number of evidences in literature. Maximum number of interactions per protein is set to 100. From the resultant PPI, the γ-quasi-clique mining algorithm described in previous section is applied to obtain any quasi-clique that overlaps the previously mined quasi-biclique on which the present human PPI has been built.

C. Human Protein-Disease Association Database

The Genetic Disease Association Database [12] (http://geneticassociationdb.nih.gov/) archives the human genetic association studies on various types of complex diseases and disorders. The database contains summary data extracted from published articles in peer reviewed journals on candidate gene and GWAS studies. The database contains both positive (if the gene/protein is known to have association with the phenotype) and negative (if a gene/protein is known to have lack of association with the phenotype) associations, and also unknown (no specific information) associations. All the gene-disease association information have been downloaded from the database and the associations other than positive ones are filtered out. The human proteins belonging to the quasi-cliques identified in the previous stage are considered and the bipartite network with these human proteins and diseases connected to them is formed. Thereafter, the γ-quasi-biclique mining algorithm is applied to this bipartite network to obtain the strong maximal quasi-bicliques from this network.

IV. Experimental Results and Discussion

A. Mining Quasi-Bicliques in HCV-Human PPIN

First we apply the proposed γ-quasi-biclique mining algorithm on the HCV-human PPIN collected from HCVpro. The value of γ has been set to 0.5. This is done as follows. We varied γ value from 0.1 to 0.9 with step size 0.1 and varied the minimum number of HCV proteins present in a quasi-biclique n from 2 to 5 with step size 1. For each combination of γ and n the algorithm is executed. In each case, the statistical significance of the set of resultant quasi-bicliques (if found) is investigated. To test the statistical significance of a quasi-biclique of size $x \times y$, the bipartite graph is perturbed randomly 10,000 times (without changing the degrees of HCV proteins) and a quasi-biclique of size $x \times y$ is picked up randomly from the perturbed graph. Then we conduct the Wilcoxon ranksum test to find whether the density of the actual quasi-biclique is significantly better than the mean density of the random quasi-bicliques of same size. This returns a p-value and lower the p-value more significant is the quasi-biclique under consideration. For a combination of γ and n value, the average p-value over all the quasi-bicliques obtained is computed and we found that for $\gamma = 0.5$ and $n = 3$ the average p-value is minimum. Hence we set the γ value to 0.5 and quasi-bicliques having at least three HCV proteins (n=3) are considered only. This results in two quasi-bicliques. Different statistics about the two quasi-bicliques found are reported in Table I. The densities (i.e., ratio of the maximum number of interactions present in the quasi-biclique to the maximum possible number of interactions) of the two quasi-bicliques obtained are 0.6786 and 0.5400, respectively. The first quasi-biclique consists of the HCV proteins CORE, NS3 and NS5A and 28 human proteins. These three HCV proteins are the top three highest degree HCV proteins in the network. The other quasi-biclique consists of five HCV proteins E1, E2, NS2, NS4A and NS5B and 10 human proteins.

B. Mining Quasi-Cliques in Human PPIN

In the next stage, the human proteins participating in the quasi-bicliques are given as the input to the STRING database. The human proteins involved in the first quasi-biclique (Table I) are first given to the STRING database with the parameter setting described in Section III-B. This induces a human interactome consisting of 120 human proteins (See Figure 1 in the Supplementary website for the interactome). After applying the quasi-clique mining algorithm described before. The γ value is fixed to 0.6 and the minimum number of nodes allowed is set to 4. This is done as per the method described in Section IV-A. From the interactome, we obtained 9 dense quasi-cliques, out of which, 5 have overlaps with the first quasi-biclique discovered in the previous stage. Different statistics of these 5 quasi-cliques are shown in Table II.

Application of the quasi-clique finding algorithm on the interactome induced by the second quasi-biclique of Table I results in 4 quasi-cliques that overlap with this quasi-biclique. The interactome induced by the second quasi-biclique consists of 79 human proteins (This interactome is shown in Figure 2 of the supplementary website). The 4 quasi-cliques are reported in Table III. As is evident, these quasi-cliques overlap with the second quasi-biclique on only one human protein each.

C. GO and Pathway Analyses of Quasi-Cliques

We further analyze the quasi-cliques found (Tables II and III) using Gene Ontology (GO) and pathway based studies based on DAVID (http://david.abcc.ncifcrf.gov/). Let us denote the 9 quasicliques identified in Table II and III by $\{QC1, QC2, \ldots, QC9\}$ respectively. Table IV shows the top few significant GO and KEGG pathway terms for the 9 quasi-cliques along with the significance p-values. It is evident from the table that for all the quasi-cliques have significant GO and KEGG pathways associated with them, with one exception for $QC7$ for which no significant KEGG pathway has been found. $QC1$ mainly consists of the proteins that function in negative regulation of ubiquitin and participate in proteasome complex whose main function is to degrade unneeded or damaged proteins by proteolysis, a chemical reaction that breaks peptide bonds. The relationship between ubiquitin, proteasome and hepatitis-c have already been reported in literature [13] which involves HCV protein CORE. It may be noticed that the HCV CORE protein belongs to the first quasi-biclique ($QB1$

TABLE I
QUASI-BICLIQUES FOUND FROM HCV-HUMAN PROTEIN INTERACTION DATABASE

Quasi-biclique	HCV proteins	Human proteins	Density
	Count: 3	Count: 28	
1	CORE, NS3, NS5A	EFEMP1, EIF2AK2, FBLN2, FBLN5, FTH1, HIVEP2, HNRNPK, JAK1, KPNA1, LTBP4, MAGED1, NAP1L1, NAP1L2, PSMB9, PSME3, RNF31, SMAD3, STAT1, STAT3, TBP, TLR2, TP53, TP53BP2, TRADD, TRAF2, TXNDC11, VIM, VWF	0.6786
	Count: 5	Count: 10	
2	E1, E2, NS2, NS4A, NS5B	CALR, CANX, CD209, CLEC4M, HOXD8, HSPA5, LTF, NR4A1, SETD2, UBQLN1	0.5400

TABLE II
QUASI-CLIQUES FOUND FROM HUMAN PROTEIN INTERACTOME OVERLAPPING WITH THE HUMAN PROTEINS OF THE FIRST QUASI-BICLIQUE (TABLE I)

Quasi-clique	Human proteins	Density	Overlapping proteins with first quasi-clique
	Count: 8		
1	POMP, PSMA2, PSMB10, PSMB7, PSMB8, PSMB9, PSME3, RFWD2	0.6786	PSMB9, PSME3
	Count: 14		
2	BIRC2, BIRC3, CASP8, FADD, GATA5, MAP3K5, RIPK1, TNFRSF1A, TNFRSF1B,, TRADD, TRAF1, TRAF2, UBC, VIM	0.6484	TRADD, TRAF2, VIM
	Count: 23		
3	CIP, EDF1, GTF2A1, GTF2A2, GTF2B, GTF2E1, GTF2F1, HNRNPK, MYST1, SETD7, SF3A2, TAF1, TAF10, TAF11, TAF12, TAF13, TAF2, TAF2E, TAF3, TAF4, TAF5, TAF7, TBP	0.6324	HNRNPK, TBP
	Count: 8		
4	HDAC1, HIPK2, MDM2, MDM4, SUMO1, TP53, UBE2I, USp7	0.6429	TP53
	Count: 8		
5	EGFR, IL6ST, JAK1, PIAS3, SRC, STAT1, STAT2, STAT3	0.7143	JAK1, STAT1, STAT3

TABLE III
QUASI-CLIQUES FOUND FROM HUMAN PROTEIN INTERACTOME OVERLAPPING WITH THE HUMAN PROTEINS OF THE SECOND QUASI-BICLIQUE (TABLE I)

Quasi-clique	Human proteins	Density	Overlapping proteins with second quasi-biclique
	Count: 4		
1	PLOD1, PLOD2, PLOD3, SETD2	0.8333	SETD2
	Count: 5		
2	NBL1, PSMD4, UBA52, UBC, UBQLN1	0.7000	UBQLN1
	Count: 45		
3	BCL2, CD3D, CREBBP, EP300, ESR1, ESR2, ESRRA, ESRRB, ESRRG, FOSB, GNG2, HNF4A, HNF4G, MAPK7, MEF2D, NFATC2, NR0B2, NR1D1, NR1D2, NR1H2, NR2C1, NR2C2, NR2C2AP, NR2E1, NR2F1, NR2F6, NR4A1, NR4A2, NR5A1, NRBP1, POMC, PPARA, PPARD, PPARG, RARA, RARB, RARG, RORA, RORB, RORC, RXRA, RXRG, THRA, THRB, VDR	0.6364	NR4A1
	Count: 5		
4	APOB, C1QA, C1QB, C1QC, CALR	0.7000	CALR

in Table I, that has overlaps with the quasi-clique $QC1$. The overlap between $QB1$ and $QC1$ consists of two human proteins PSMB9 and PSME3 and thus they may be considered as possible infection gateway by the HCV proteins CORE (interacts with PSME3), NS3 (interacts with PSMB9) and NS5A (interacts with PSMB9) which belong to quasi-biclique $QB1$, for attacking the proteasome complex.

The quasi-clique $QC2$ contains 14 proteins mostly involved in apoptosis and programmed cell death. Also a significant GO-CC term for these proteins is death-inducing signaling complex. These proteins also participate in the KEGG pathway apoptosis as well as pathways in cancer. These evidences suggest that the human proteins involved in $QC2$ have relationships with the cancer diseases. The quasi-biclique $QB1$ (involving the viral proteins CORE, NS5A and NS3) overlaps with $QC2$ on three human proteins TRADD (interacts with CORE and NS5A), TRAF2 (interacts with CORE and NS5A) and VIM (interacts with CORE and NS3). This suggests that attack by HCV proteins CORE, NS5A and NS3 may lead to cancer through apoptosis and the main gateway host proteins responsible for that are TRADD, TRAF2 and VIM.

The 23 host proteins in quasi-clique $QC3$ are mainly transcription factors (Table IV). Although the quasi-biclique $QB1$ only overlaps with $QC3$ on two host proteins HNRNPK and TBP, it suggests that the viral proteins in $QB1$ may indirectly interact with many transcription factor proteins and thus may cause their malfunctioning. This may lead to breakdown of the overall setup of normal regulatory roles of these transcription factors causing serious infectious behavior.

Most of the host proteins in $QC4$ negatively regulate transcription and participate in enzyme binding. Moreover, many of these proteins are part of PML bodies, which is a class of nuclear body and they react against SP100 autoantibodies (PML, promyelocytic leukemia). Also pathway analysis finds two significant KEGG pathways, namely p53 signaling pathway and chronic myeloid leukemia. For the quasi-biclique $QB1$ the viral gateway to these host proteins is TP53, a membrane protein that is common for $QB1$ and $QC4$. Noticeably, all the viral proteins of $QB1$, i.e., CORE, NS5A and NS3 interact with TP53 to get entrance. This infection may ultimately lead to chronic myeloid leukemia [14].

The quasi-clique $QC5$ contains host proteins with mainly kinase activities. Two significant KEGG pathways namely JAK-STAT signaling pathway and pancreatic cancer, are identified in this quasi-clique. This suggests that the HCV proteins in $QB1$ interact with the host proteins in $QC5$ through

the common host proteins JAK1, STAT1 and STAT3 leading to pancreatic cancer. Moreover, JAK-STAT system transmits information from chemical signals outside the cell, through the cell membrane. Therefore the proteins involved in $QC5$ are possibly involved in transferring and propagating the infection to the other cells. A study in [15] has already established the involvement of HCV in JAK-STAT signaling pathway.

The quasi-cliques $QC6$ through $QC9$ (Table III) overlap with the quasi-biclique $QB2$, which consists of 5 viral proteins E1, E2, NS2, NS4A, and NS5B and 10 host proteins. $QB2$ overlaps with $QC6$ with the host protein SETD2. The most significant GO terms associated with the human proteins in $QC6$ in BP, MF and CC categories are oxidation reduction, procollagen-lysine 5-dioxygenase activity and endoplasmic reticulum, respectively. The most significant KEGG pathway associated with these proteins is Lysine degradation, where all the 4 proteins in $QC6$ are involved. The association of HCV NS2 protein and lysine degradation is also reported in [16].

$QC7$ overlaps with $QB2$ with the host protein UBQLN1. $QC7$ also has proteasomal acitivities $QC1$, and the host proteins in this functional module are involved in HCV infection. However, we could not find any significant pathway for $QC7$.

$QC8$ is the largest quasi-clique consisting of 45 host proteins which are mostly transcription factors. The infection gateway to this module is NR4A1, which is the only common host protein for $QB2$ and $QC8$. Interestingly, all the five viral proteins in $QB2$ interact with NR4A1, and the CORE protein, which is a part of $QB1$ also interacts with NR4A1. This observation suggests that NR4A1 serves as a very important gateway to this transcription factor complex. Any disturbance to this module for viral infection may lead to malfunctioning of normal gene regulatory network, and this in turn can result in various types of cancer (as the pathway study reveals). Our pathway study also reveals another significant pathway, namely PRAR signaling pathway, which is also shown to be associated with HCV infection in recent studies [17].

$QC9$ consists of 5 host proteins associated with protein maturation and humoral immune response mediated by circulating immunoglobulin. Thus these proteins are highly responsible for maintaining the immunity system inside human body. $QB2$ and $QC9$ have one common host protein CALR, and hence this protein serves as a gateway of attack to the immunity system by HCV. The viral proteins E1 and E2 (envelop proteins), which are major players in all events required for virus entry into target cells, interact with CALR and start attacking the immunity system. This may ultimately lead to many prion diseases (as revealed through pathway analysis).

The GO and pathway analyses of the quasi-cliques in human PPIN reveal that the host proteins involved in these functional modules have high degree of similarities. Moreover, HCV attacks that go through these quasi-cliques may lead malfunctioning of regulatory and immunity system in targeted cells leading to different types of disease including cancers.

D. Mining Quasi-Bicliques in Human Protein-Disease Association Network

Next, we apply our quasi-biclique finding algorithm on the human gene-disease association network. Note that while finding the quasi-bicliques, we executed the quasi-biclique finding method on 9 different bipartite graphs, corresponding to the 9 quasi-cliques. Each of these graphs contain the human proteins from the corresponding quasi-clique, and all the diseases. The γ value is set to 0.7. This is done using the method described in Section IV-A. Out of 9 quasi-cliques, we found 4 quasi-cliques $QC1$, $QC2$, $QC4$ and $QC8$ which have overlap with the obtained quasi-bicliques on protein-disease association networks. These quasi-bicliques, termed as $QBD1, QBD2, QBD3, QBD4$ are reported in Table V. In each quasi-biclique in human protein-disease association network, two human proteins have been found to overlap with the corresponding quasi-cliques. These proteins can be considered as gateways to the diseases. $QBD1$ has overlap with $QC1$ with two proteins PSMB8 and PSMB9 which are associated with five different diseases. $QBD2$ overlaps with $QC2$ with two host proteins TNFRSF1A and TNFRSF1B that are highly associated with 12 diseases. The quasi-clique $QC4$ and the quasi-biclique $QBD3$ have two common proteins $TP53$ and $MDM2$ which are connected two 9 diseases including various types of cancer. Two proteins TGFR and MDM2 are common to $QBD4$ and $QC8$ and these proteins have association with 5 diseases (mainly different cancer types). Interestingly MDM2 belongs to both $QBD3$ and $QBD4$.

As is evident from Table V, several diseases are associated with the 4 quasi-bicliques in human protein-disease association network. Among these, many of the diseases are already established to be related to HCV infection. Graves' disease is an autoimmune disease where the thyroid is overactive. It has been found recently that chronic HCV infection may lead to destructive thyroiditis followed by Graves' disease [18]. Diabetes (Type I and II) is a well-known disease to be associated with HCV attack [19]. Interferons are proteins that are released during the presence of viral particles in cells. It has been established recently that HCV infection suppresses the interferon response in the liver [20]. The relationship of Psoriasis, another autoimmune disease affecting skin, is also well-known [21]. We have also found malaria as one of the diseases in the quasi-bicliques. A recent study has revealed that HCV infection may lead to slower emergence of malaria parasite *Plasmodium falciparum* in blood [22]. Chron's disease is the condition of continuous inflammation of digestive track. Inflamamtory bowel diseases (IBD) such as Chron's disease or colitis are established to be linked with viral hepatitis [23]. Also systemic lupus erythematosus has been found to be more prevalent in HCV infected patients [24]. Rheumatoid Arthritis, a common disease inducing inflammation in joints is also well-linked with HCV infection and people with HCV often show raised levels of rheumatoid factor in their blood [25]. Table V also reports some types of cancer to be associated with the proteins in the quasi-bicliques. Recent research has focused on development of cancer in HCV infected patients and different studies have established the links between hepatitis-C and various types of cancers such as liver cancer [26], breast cancer [27], leukemia [28], colorectal cancer [29], endometrial cancer [29], and lung cancer [30]. As depicted in the table, HCV infection has also been found to be associated with a higher risk of coronary diseases [31]. As many of the diseases

TABLE IV
THE SIGNIFICANT IMPORTANT GO TERMS AND KEGG PATHWAYS FOUND IN THE QUASI-CLIQUES

Quasi-clique	Significant GO terms			KEGG Pathway
	Biological Process	Molecular Function	Cellular Component	
$QC1$	negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle (p-value: 4.6e-11, 75%)	threonine-type endopeptidase activity (p-value: 6.1e-11, 62.5%)	proteasome complex (p-value: 6.4e-14, 87.5%)	Proteasome (p-value: 3.1e-12, 87.5%)
$QC2$	apoptosis (p-value: 8.9e-14, 85.7%) programmed cell death (p-value: 1.1e-13, 85.7%)	death domain binding (p-value: 5.0e-3, 14.3%)	membrane raft (p-value: 3.9e-8, 42.9%) death-inducing signaling complex (p-value: 5.5e-6, 21.4%)	Apoptosis (p-value: 1.1e-10, 57.1%) pathways in cancer (p-value: 3.6e-4, 42.9%)
$QC3$	transcription initiation from RNA polymerase II promoter (p-value: 6.0e-29, 71.4%)	general RNA polymerase II transcription factor activity (p-value: 4.9e-20, 52.4%)	DNA-directed RNA polymerase II, holoenzyme (p-value: 4.1e-30, 76.2%)	Basal transcription factors (p-value: 1.8e-29, 71.4%)
$QC4$	negative regulation of transcription (p-value: 1.0e-8, 87.5%)	enzyme binding (p-value: 1.2e-3, 50.0%,)	PML body (p-value: 1.6e-7, 50.0%)	p53 signaling pathway (p-value: 1.0e-3, 37.5%) Chronic myeloid leukemia (p-value: 1.3e-3, 37.5%)
$QC5$	protein kinase cascade (p-value: 2.8e-9, 87.5%)	protein tyrosine kinase activity (p-value: 3.3e-3, 37.5%)	dendrite (p-value: 7.4e-2, 25.0%)	Jak-STAT signaling pathway (p-value: 4.9e-7, 75.0%) Pancreatic cancer (p-value: 9.1e-5, 50.0%)
$QC6$	oxidation reduction (p-value: 1.0e-4, 100.0%)	procollagen-lysine 5-dioxygenase activity (p-value: 1.1e-7, 75.0%)	endoplasmic reticulum (p-value: 5.6e-3, 75.0%)	Lysine degradation (p-value: 6.0e-7, 100.0%)
$QC7$	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process (p-value: 2.3e-5, 60.0%)	structural constituent of ribosome (p-value: 2.6e-2, 40.0%)	cytosolic small ribosomal subunit (p-value: 1.2e-2, 40.0%) proteasome complex (p-value: 1.9e-2, 40.0%)	–
$QC8$	regulation of transcription, DNA-dependent (p-value: 4.3e-27, 84.4%)	steroid hormone receptor activity (p-value: 6.1e-75, 73.3%) transcription factor activity (p-value: 7.7e-37, 84.4%)	nuclear lumen (p-value: 1.4e-3, 20.0%) transcription factor complex (p-value: 4.7e-3, 8.9%)	Pathways in cancer (p-value: 1.7e-5, 20.0%) PPAR signaling pathway (p-value: 1.3e-4, 11.1%)
$QC9$	protein maturation (p-value: 2.8e-6, 80.0%) humoral immune response mediated by circulating immunoglobulin (p-value: 3.0e-5, 60.0%)	carbohydrate binding (p-value: 5.4e-2, 40.0%)	extracellular space (p-value: 5.9e-4, 80.0%)	Prion diseases (p-value: 1.4e-4, 60.0%) Complement and coagulation cascades (p-value: 5.4e-4, 60.0%)

reported in our study already have evidence in literature for their association with HCV infection, the quasi-cliques and quasi-bicliques obtained here may put light on the possible pathways of HCV pathogenesis leading to these diseases.

E. Analyses of Gateway Proteins

Previous results and discussions have pointed out two types of gateway proteins, one set acts as the gateway to the host cellular mechanism for the viral proteins, and the second set consists of the host proteins that have high degree of association to different kinds of diseases. The first set VH (Viral-Host) contains 15 host proteins: PSME3, TP53, TBP, TRADD, STAT3, HNRNPK, NR4A1, SETD2, PSMB9, TRAF2, STAT1, CALR, JAK1, VIM and UBQLN1 (Tables II and III). The second set HD (Human-Disease) contains 7 host proteins PSMB8, PSMB9, TNFRSF1A, TNFRSF1B, TP53, MDM2 and EGFR. The results reveal that HCV infection pathogenesis should propagate through the proteins in VH and HD sets, and thus these proteins play extremely important role during viral infection. Specially, the proteins in the set VH are responsible for the initiation of the infection process. First we compare the average degrees of gateway and non-gateway proteins and found that average degree of gateway proteins is 21.6364, whereas the average degree of non-gateway proteins is 4.2295. The difference is statistically significant as per Wilcoxon's rank sum test (p-value: 1.3006e-09). This suggests that the viral proteins tend to attack high-degree host proteins for initiating infection. Moreover, to test whether these proteins have some unique features, we investigate for their GO (BP) and pathway enrichment (Table VI). It is evident that the significant GO-BP terms mostly involved in apoptosis and programmed cell death which indicates that the targeted host proteins are highly associated with the process of cell death. Moreover significant pathways suggest that HCV infection ultimately lead to various cancer types including pancreatic cancer which is already established in a recent study [32].

V. CONCLUSIONS

In this article a system-wide study has been made for identifying possible infection pathway of HCV. For this purpose, quasi-bicliques in HCV-human PPIN are mapped onto quasi-cliques in human PPIN. Subsequently, the quasi-cliques are mapped onto human protein-disease association networks. Quasi-clique and quasi-biclique mining algorithms have been proposed in this context. The quasi-cliques that overlap with the quasi-bicliques in HCV-human PPIN have been found to contain host proteins highly associated in various disease pathways including different cancer types. Many of the diseases have evidence in literature for their connection with HCV infection. Further, the gateway proteins, i.e., the proteins which are mainly targeted by HCV proteins to disturb the host cellular mechanisms, have been found to have high degrees in human interactome compared to the other virus-targeted proteins. Moreover, the gateway proteins are tested for GO-BP enrichment and pathway enrichment, and these analyses

TABLE V
QUASI-BICLIQUES FOUND FOR HUMAN PROTEIN-DISEASE ASSOCIATION NETWORK CORRESPONDING TO FOUR QUASI-CLIQUES

Quasi-biclique	Corresponding QC	Human proteins	Diseases	Density
		Count: 2	Count: 5	
$QBD1$	$QC1$	PSMB8, PSMB9	Graves disease, diabetes (type 1), interferon response, psoriasis, malaria; hypoglycemia; hyperparasitemia	0.7000
		Count: 2	Count: 12	
$QBD2$	$QC2$	TNFRSF1A, TNFRSF1B	Crohn's disease, ulcerative colitis, cystic fibrosis, Lupus, Rheumatoid Arthritis, diabetes (type 2), amyloidosis, breast cancer, Tumor necrosis factor receptor-associated periodic syndrome, bone density, bone mass	0.7083
		Count: 2	Count: 9	
$QBD3$	$QC4$	TP53, MDM2	DNA Damage|Lung Neoplasms, B-Cell Chronic Lymphocytic Leukemia, bladder cancer, breast cancer, colorectal cancer, endometrial cancer, liver cancer, lung cancer, stomach cancer	1.000
		Count: 2	Count: 5	
$QBD4$	$QC8$	EGFR, MDM2	colorectal cancer, lung cancer, Acute Coronary Syndrome, Breast Neoplasms Carcinoma| Non-Small-Cell Lung|Exanthema|Lung Neoplasms	0.7000

TABLE VI
SIGNIFICANT GO-BP AND KEGG PATHWAY TERMS FOR VIRAL-HUMAN GATEWAY PROTEINS

Significant GO-BP terms
cytokine-mediated signaling pathway (p-value: 3.7e-5)
regulation of apoptosis (p-value: 5.2e-5)
regulation of programmed cell death (p-value: 5.5e-5)
regulation of cell death (p-value: 5.6e-5)
positive regulation of macromolecule metabolic process (p-value: 7.4e-5)
Significant KEGG pathways
Pancreatic cancer (p-value: 5.5e-4)
Pathways in cancer (p-value: 5.6e-3)

reveal that these proteins are highly involved in apoptosis and programmed cell death leading to various cancer types.

REFERENCES

[1] A. Panchenko and T. Przytycka, *Protein-protein Interactions and Networks: Identification, Computer Analysis, and Prediction*, vol. 9. London: Springer-Verlag, 2008.

[2] M. R. Arkin and J. A. Wells, "Small-molecule inhibitors of protein-protein interactions: progressing towards the dream.," *Nature Reviews Drug Discovery*, vol. 3, no. 4, pp. 301–317, 2004.

[3] O. Tastan et. al., "Prediction of interactions between HIV-1 and human proteins by information integration," in *Proc. Pacific Symp. Biocomputing*, pp. 516–527, 2009.

[4] J. I. MacPherson et. al., "Patterns of HIV-1 Protein Interaction Identify Perturbed Host-Cellular Subsystems," *PLoS Comput Biol*, vol. 6, no. 7, pp. e1000863+, 2010.

[5] A. Mukhopadhyay et. al., "Mining association rules from hiv-human protein interactions," in *Proc. ICSMB 2010*, pp. 344–348, 2010.

[6] U. Maulik et. al., "Identifying the immunodeficiency gateway proteins in humans and their involvement in microrna regulation," *Mol. BioSyst.*, vol. 7, no. 6, pp. 1842–1851, 2011.

[7] A. Mukhopadhyay et. al., "A novel biclustering approach to association rule mining for predicting hiv-1–human protein interactions," *PLoS ONE*, vol. 7, no. 4, p. e32289, 2012.

[8] I. C. Lorenz, "The hepatitis C virus nonstructural protein 2 (NS2): An up-and-coming antiviral drug target," *Viruses*, vol. 2, no. 8, pp. 1635–1646, 2010.

[9] C.-I. Popescu et. al., "NS2 protein of hepatitis C virus interacts with structural and non-structural proteins towards virus assembly," *PLoS Pathog*, vol. 7, no. 2, p. e1001278, 2011.

[10] S. K. Kwofie et. al., "HCVpro: Hepatitis C virus protein interaction database," *Infect Genet Evol*, September 2011.

[11] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[12] Y. Zhang et. al., "Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information," *BMC Med. Genomics*, vol. 21, no. 31, 2010.

[13] R. Suzuki et. al., "Proteasomal turnover of hepatitis C virus core protein is regulated by two distinct mechanisms: a ubiquitin-dependent mechanism and a ubiquitin-independent but PA28gamma-dependent mechanism," *J. Virology*, vol. 83, no. 5, pp. 2389–2392, 2009.

[14] J. M. Klco et. al., "Bone marrow biopsy in patients with hepatitis C virus infection: spectrum of findings and diagnostic utility," *American J. Hematology*, vol. 85, no. 2, pp. 106–110, 2010.

[15] L. Zhang et. al., "IL28B inhibits hepatitis C virus replication through the JAK-STAT pathway," *J Hepatol.*, vol. 55, no. 2, pp. 289–298, 2011.

[16] S. Welbourn et. al., "Investigation of a role for lysine residues in non-structural proteins 2 and 2/3 of the hepatitis C virus for their degradation and virus assembly," *J Gen Virol.*, vol. 90, no. 5, pp. 1071–1080, 2009.

[17] F. Negro, "Peroxisome Proliferator-Activated Receptors and Hepatitis C virus-induced insulin resistance," *PRAR Research*, 2009.

[18] B. K. Kim et. al., "Interferon-alpha-induced destructive thyroiditis followed by Graves' disease in a patient with chronic hepatitis C: a case report," *J. Korean Med. Sci.*, vol. 26, no. 12, pp. 1638–1641, 2011.

[19] C. Giordanino et. al., "Type 2 diabetes mellitus and chronic hepatitis C: which is worse? Results of a long-term retrospective cohort study," *Dig Liver Dis*, vol. 44, no. 5, pp. 406–412, 2012.

[20] M. Tsuge et. al., "Hepatitis C virus infection suppresses the interferon response in the liver of the human hepatocyte chimeric mouse," *PLoS ONE*, vol. 6, no. 8, p. e23856, 2011.

[21] T. Yamamoto et. al., "Psoriasis and hepatitis C virus," *Acta Derm. Venereol.*, vol. 75, no. 6, pp. 482–483, 1995.

[22] O. Ouwe-Missi-Oukem-Boyer et. al., "Hepatitis C virus infection may lead to slower emergence of P. falciparum in blood," *PLoS ONE*, vol. 6, no. 1, p. e16034, 2011.

[23] J. K. Hou et. al., "Viral hepatitis and inflammatory bowel disease," *Inflamm. Bowel Dis.*, vol. 16, no. 6, pp. 925–932, 2010.

[24] G. Perlemuter et. al., "Hepatitis C virus infection in systemic lupus erythematosus: a case-control study," *J. Rheumatol.*, vol. 30, no. 7, pp. 1473–1478, 2003.

[25] C. Ferri et. al., "Current treatment of hepatitis C-associated rheumatic diseases," *Arthritis Res Ther*, vol. 14, no. 3, p. 215, 2012.

[26] L. J. de Oliveria Andrade et. al., "Association between hepatitis C and hepatocellular carcinoma," *J Glob Infect Dis*, vol. 1, no. 1, pp. 33–37, 2009.

[27] F. H. Su et. al., "Association between chronic viral hepatitis infection and breast cancer risk: a nationwide population-based case-control study," *BMC Cancer*, vol. 11, p. 495, 2011.

[28] G. Emilia et. al., "Hepatitis C virus-induced leuko-thrombocytopenia and haemolysis," *J. Med. Virol.*, vol. 53, no. 2, pp. 182–184, 1997.

[29] A. A. Mohamed et. al., "Chronic hepatitis c genotype-4 infection: role of insulin resistance in hepatocellular carcinoma," *Virol. J.*, vol. 8, p. 496, 2011.

[30] L. H. Omland et. al., "Hepatitis C virus infection and risk of cancer: a population-based cohort study," *Clin Epidemiol*, vol. 2, pp. 179–186, 2010.

[31] A. A. Butt et. al., "Hepatitis C virus infection and the risk of coronary disease," *Clin. Infect. Dis.*, vol. 49, no. 2, pp. 225–232, 2009.

[32] H. B. El-Serag et. al., "Risk of hepatobiliary and pancreatic cancers after hepatitis c virus infection: A population-based study of u.s. veterans," *Hepatol.*, vol. 49, no. 1, pp. 116–123, 2009.