Novel Biomarkers Discovery for HBV and HCV Monitoring Through Protein Interaction Networks Analysis

Thomas Simos, Costas Papaloukas, George Thyphronitis and Urania Georgopoulou

***Abstract*—According to the World Health Organization hepatitis is a global threat to public health. Various statistics reveal that hundreds of millions of people worldwide are infected by chronic hepatitis C (HCV), which accounts for only the 15% of viral hepatitis. Hepatitis B (HBV) is the second chronic type of the virus with even larger numbers that reach up to 350 million carriers. Several research efforts have been focused recently on the underlying mechanisms of the infection and particularly on the assessment of gene/protein expressions. The utilization of protein interaction networks (PINs) is expected to identify novel aspects of the disease concerning both the patients' immune response and preventive treatment regimens. Here we designed several PINs for HCV and HBV and employed topological, modular and functional analysis techniques in order to determine significant network nodes that correspond to prominent candidate biomarkers.**

I. Introduction

Hepatitis is the condition of liver inflammation that may be caused by a virus, drugs, alcohol or any other factor that affects the cells of the liver, the hepatocytes. However hepatitis is most commonly caused by a virus. Hepatitis C virus (HCV) infection affects 170 million of individuals worldwide and has a high rate of chronicity. Patients with chronic infection present liver injury, caused by immune mechanisms and metabolic disorders related to hepatic fibrogenesis, steatosis and insulin resistance. These patients have high probability to develop liver cirrhosis and cancer [1]. Even today the molecular mechanisms of HCV pathology remain partially understood. The 9.6 kb HCV genome encodes a polyprotein that translates four structural proteins, namely CORE, E1, E2 and p7, and seven nonstructural, namely NS2, NS3, NS4A, NS4B, NS5A, NS5B and F.

Hepatitis B is one of the most common infectious diseases in the world and almost 50 years after its discovery, it still has a major impact on health with more than 350 million chronic carriers worldwide [2]. The hepatitis B virus (HBV) genome is around 3.2 kb in length and composed of four overlapping open reading frames that cover the entire genome and encode eight proteins (small S, middle S, S, pre-Core, Core, HBV Polymerase, P and HBx). The HBV infection initially causes an asymptomatic short phase but eventually will either be expelled by the organism or may lead to chronic infection.

The diagnosis is made by detecting the surface antigen of hepatitis B virus (HBsAg) in the blood. Other biomarkers are also examined, like the HBeAg antigen, the antibody e, the antibody S and the core antibody [3]. Similarly, the serological assays that detect the antibody to HCV contain antigens from the CORE and the NS3-NS5 genes. These procedures are both expensive and time consuming and cannot be used to monitor efficiently the patients' condition or the progress of the disease.

On the other hand, cellular functions are coordinated activities of many proteins and biological molecules that interact with each other. A system of interacting elements can be abstracted with the mathematical structure of a graph. In most studies of biological networks, system elements like proteins and genes are depicted by graph nodes and the physical interactions between those elements by edges. Protein interaction networks (PINs) are commonly represented by undirected graphs and can be characterized by several properties [4].

Our aim is to identify novel biomarkers for monitoring both HCV and HBV, which can be detected and measured with a simple blood test. For that, we designed the PINs for HCV and HBV using protein interactions gathered from several related biomedical databases (DBs). Then we evaluated their topological, modular and functional properties and particularly the role of hub proteins (i.e. highly connected proteins). The derived data were assessed in order to determine key properties among them and discover the best candidate biomarkers.

II. Materials and Methods

A. Protein Interaction Data

In order to construct the employed PINs for HBV and HCV analysis we used data from various relevant DBs, namely BOND, IntAct, VirusMINT, VirHostNet, HCVpro, BioGRID, DIP, HPRD and Reactome. The data from each DB exhibit certain characteristics that provide complementary information to the design of our PINs. In particular, BOND is the Biomolecular Object Network Databank which contains a variety of DBs, including

Manuscript received July 30, 2013. This research project has been co-financed by the European Union (European Regional Development Fund-ERDF) and Greek national funds through the Operational Program "THESSALY- MAINLAND GREECE AND EPIRUS-2007-2013" of the National Strategic Reference Framework (NSRF 2007-2013).

Th. Simos is with the Department of Biological Applications and Technology, University of Ioannina, 45110 Ioannina, Greece (e-mail: thsimos@cc.uoi.gr).

C. Papaloukas is with the Department of Biological Applications and Technology, University of Ioannina, 45110 Ioannina, Greece (corresponding author; phone: +30-26510-07427; e-mail: papalouk@cc.uoi.gr).

G. Thyphronitis is with the Department of Biological Applications and Technology, University of Ioannina, 45110 Ioannina, Greece (e-mail: gthyfron@uoi.gr).

U. Georgopoulou is with the Molecular Virology Laboratory, Hellenic Pasteur Institute, 11527 Athens, Greece (e-mail: uraniag@pasteur.gr).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

GenBank and BIND. The Biomolecular Interaction Network Database (BIND) is a collection of records documenting over 175000 molecular interactions for nearly 3000 protein complexes and several pathways, drawn from publications and high-throughput experiments [5].

IntAct is an open data molecular database with interactions either from the literature or from direct data depositions. It contains 275000 verified interactions attained from more than 5000 publications [6].

In VirusMINT DB all interactions between viral and human proteins are collected from the literature [7]. The DB contains over 5000 interactions involving more than 490 viral proteins from more than 110 different viral strains.

VirHostNet is dedicated to the development and analysis of PINs between viruses and human. It contains data from taxonomy, interactome, networks and text-mining. The human interactome, in particular, is currently composed of 72357 interactions. The DB provides also data for 5175 more interactions between 1474 cellular proteins and 1162 viral proteins from 220 viral strains [8].

The HCV Protein Interaction DB (HCVpro) is a specially tailored knowledge-base for HCV protein interactions. It contains 621 manually verified literature and DB curated interactions between HCV and host human cellular proteins. HCVpro includes canonical pathways, gene ontologies and microarray expression data that can facilitate the discovery of drugs, drug targets and diagnostic biomarkers [9].

BioGRID is a public DB that archives and disseminates genetic and protein interaction data from model organisms and humans. It holds over 630000 interactions curated from both high-throughput datasets and individual focused studies, as derived from over 37000 publications in the primary literature [10].

The Database of Interacting Proteins (DIP) stores data on experimentally determined protein-protein interactions. It combines information from a variety of sources to create a single consistent set of records. The DB provides data for 75420 interactions between 25628 proteins [11].

The Human Protein Reference Database (HPRD) is an object oriented platform for the integration and visualization of information related to domain architecture, post-translational modifications, interaction networks and disease associations for the proteins in the human proteome. It contains 41327 protein-protein interactions for 30047 protein entries [12].

Finally, Reactome is a curated DB of pathways and reactions (pathway steps) in human biology. In Reactome the definition of a 'reaction' includes binding, activation, translocation and degradation, in addition to the classical biochemical reactions. It provides data for 6478 reactions between 6981 human proteins [13].

B. Network Analysis

We searched the above DBs with the 11 HCV and the eight HBV proteins, separately for each type of virus, to find interactions between them and the human proteins. The derived protein interactions were imported to Cytoscape. Cytoscape is an open source platform for complex network analysis and visualization. It can be used to design molecular interaction networks or biological pathways and integrate them with annotations or other available data [14].

For each database we created separate network visualizations. Subsequently the networks were merged for each type of hepatitis into two separate overall PINs (HCV and HBV). To avoid any duplicate entries between the DBs due to different synonyms, we used the Entrez symbols of the human genes that encode the interacting cellular proteins as well as their assigned UniProt IDs.

For the topological analysis of both networks we calculated several topological parameters such as the number of nodes and edges, the clustering coefficient, the connected components, the network diameter and centralization, the characteristic path length, the average number of neighbors, the node degree distribution and the neighborhood connectivity distribution (Table I) [15]. The majority of these parameters refers to the general architecture of a network and provides overall statistics about its size, density and connectivity. The more complex parameters estimate certain distributions in the network that point out the distinct characteristics of its topology.

In specific, *degree* is a topological index that corresponds to the number of nodes directly connected to a given node. Based on the degree calculation we can define the degree distribution $P(k)$, which estimates the probability of a node to have exactly k links. A network following a power law degree distribution, i.e. $P(k) = ak^{-\gamma}$ (where a is a scaling factor and γ a positive constant), indicates that there are many low-degree and few high-degree nodes in this network. Nodes with high degree (highly connected) are called *hubs* and hold together several nodes of lower degree. Likewise, the *neighborhood connectivity* distribution gives the average of the neighborhood connectivity of all nodes with k neighbors ($k = 0,1, ...$). If the distribution decreases then most edges in the network connect low degree nodes with high degree nodes, which is an indicator that the network consists of subnetworks.

In addition to the topological features described above we investigated also the modular aspects of the two PINs. Modularity analysis detects protein complexes and functional modules within the interaction networks [16]. A protein complex is a group of proteins that interact with each other and form a multi-molecular subsystem. The difference with a functional module is that in the latter the protein cluster is involved in a particular cellular process.

In modularity analysis the basic elements are *cliques* and *cores*. A *clique* defines a subnetwork within the network that is complete, i.e. it contains nodes that are fully connected to each other. A *maximum clique* is the largest clique in the network while a *maximal clique* is a clique not contained to any other clique. A *k-core* is a subnetwork where all nodes are connected to at least k other nodes within the core. Core estimation can be used to detect subnetworks of certain density in large networks.

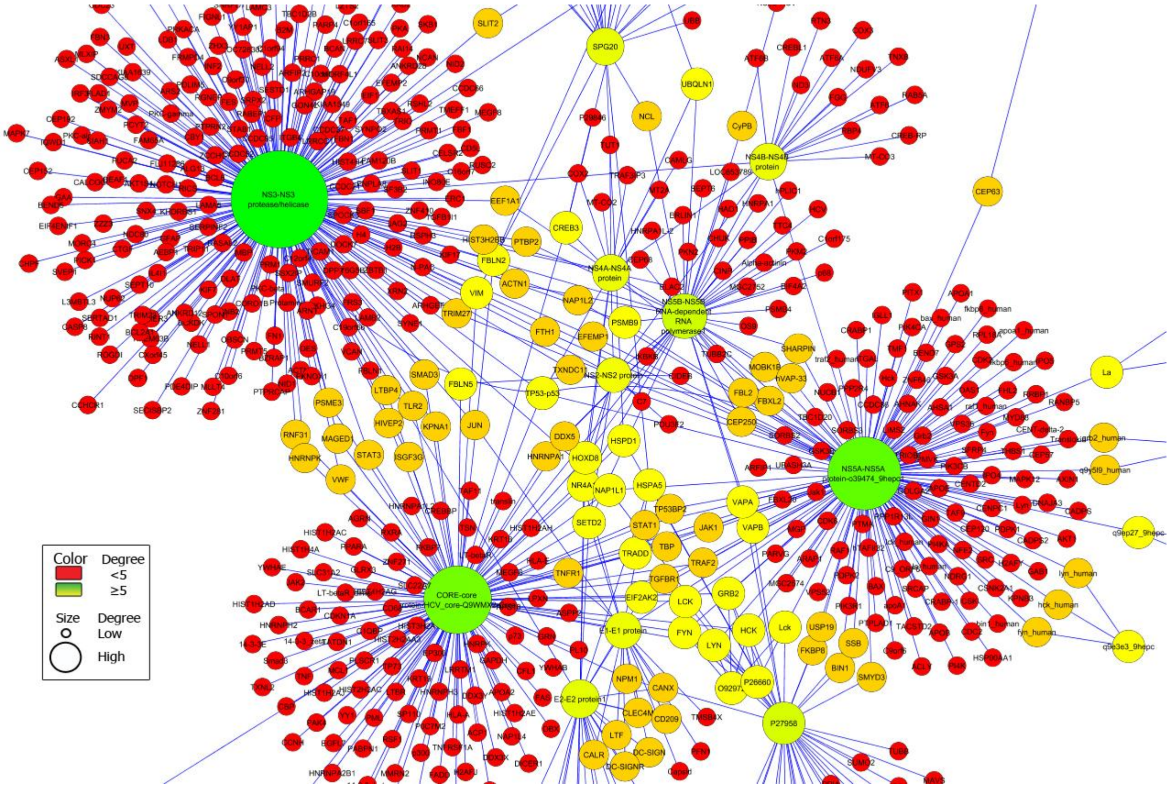

Fig. 1: Main part of the HCV overall Protein Interaction Network. A proteome-wide mapping approach of interactions between HCV and cellular proteins was performed to provide a comprehensive view of viral infection.

C. Biomarkers Discovery

In order to identify intrinsic nodes to our PINs, we combined topological, modular and functional analysis tools that produced a set of features for each protein node. Specifically we estimated their degree, betweeness and maximal clique centralities in order to determine the importance of each node in the networks. The betweeness centrality of node n is calculated by dividing the number of shortest paths that connect v_i and v_j and pass through node n with the total number of shortest paths that connect v_i and v_j. This measure reveals the nodes that control the interactions of other nodes in the network. Similar to the betweeness centrality is the maximal clique centrality, which also identifies essential nodes in the network but through modular analysis and the utilization of the clique components. As for functional analysis, it aims at the enrichment of the networks with semantic terms from the Gene Ontology (GO) database and specifically the use of similarity analysis for calculating the degree and betweeness centrality based on GO data. Semantic similarities here were used as a means of incorporating functional information about the nodes to the network and increase that way the reliability of their interactions.

III. Results and Discussion

A. PINs Analysis

The HCV network consists of 903 interactions observed between 814 human and virus proteins, while the HBV network consists of 275 interactions and 253 proteins. Figure 1 depicts the visualization of the HCV network based on the number of interactions (degree) of each protein. The size of each node depends on the number of nodes that interacts with. The red color indicates that these proteins have up to four interactions and for more than 5 the color gradually varies from yellow to green as the number of interactions increases. Here we denote proteins with a degree of five or more as hubs and are considered to have a significant role in the network. This degree centrality criterion has been used to characterize the importance of a node in several other studies as well [17].

TABLE I
Topological Parameters of the Protein Interaction Networks for the two Types of Hepatitis

Parameters	**HBV**	**HCV**
Number of nodes	253	814
Number of edges	275	903
Clustering coefficient	0	0.052
Connected components	5	48
Network diameter	11	10
Network centralization	0.347	0.299
Characteristic path length	3.793	3.368
Average number of neighbors	2.174	2.219
Node degree distribution (γ)	0.928	0.982
Neighborhood Connectivity Distribution (γ)	0.966	0.685

Table I depicts the main topological parameters for the HBV and HCV networks. Other features were also assessed but those included in the table describe the prominent architectural properties for the two PINs. The values of the topological parameters for HBV and HCV are relatively close, indicating that both PINs have similar architecture.

The node degree distribution of both networks is consistent with the power law and characterizes them as *scale free*. This is depicted from the value of γ, which ranges from 0.928 to 0.982. Both PINs have few high degree hub proteins, while most nodes have few connections, as illustrated in Figure 2.

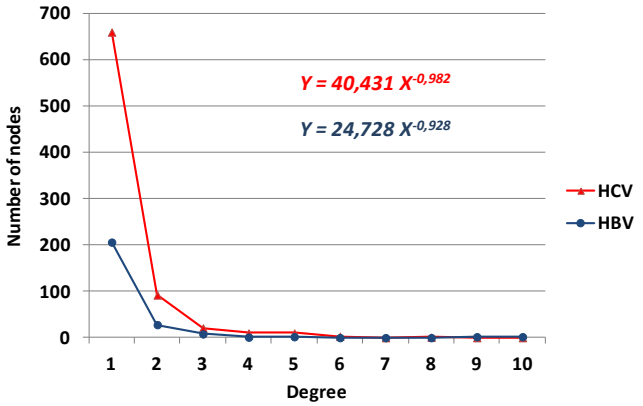

Fig. 2: Node degree distribution of HBV and HCV PINs. The approximations to the corresponding power law equations are also given.

B. Novel Biomarkers

As a first approach we tried to normalize the values for the employed topological, modular and functional features and sum up their values in order to classify the protein nodes. But following the notion of gene ranking used in microarray data analysis [18], [19] we estimated instead the deviation of each feature from its mean value, normalized it and then sorted the summed values. In this way more prominent conclusions can be drawn since each feature value has smaller effect to the statistical ranking. From the sorted list we selected only the proteins that refer to nodes that are common in both HBV and HCV PINs. This yielded a list of 42 proteins. From this list we subsequently considered only those with a positive overall value thus only four proteins remained, namely HSPA5, STAT3, TP53 and MIF. It should be noted that some of the markers previously proposed in the literature for the prognosis and treatment of hepatitis were scored high in the derived sorted list (e.g. ALB and CCR5) while others much worse than the ones proposed here (e.g. APOA1, TNF and FAS). Moreover, proteins STAT3 and TP53 have been quite recently associated with hepatitis [20], [21] indicating strongly the robustness of the proposed approach.

IV. Conclusion

The obtained results indicate that both PINs have similar overall architecture. The values of almost all topological features are very close except those referring to the networks' connectivity (nodes and edges) which is mostly due to the more extensive study of HCV. The intersection of the two PINs revealed 42 common protein nodes [22], some of which are already used as biomarkers. By applying a statistical ranking procedure we identified four of them as of higher significance. Evidently, our next step is to validate their classification efficacy under clinical conditions.

V. References

[1] B. de Chassey *et al.*, "Hepatitis C virus infection protein network," *Mol. Syst. Biol.*, vol. 4, article 230, 2008.

[2] J. H. Kao and D. S Chen, "Global control of hepatitis B virus infection," *Lancet Infect. Dis.*, vol. 2, pp.395–403, 2002.

[3] P. Dény and F. Zoulim, "Hepatitis B virus: From diagnosis to treatment," *Pathologie Biologie*, vol. 58, pp. 245–253, 2010.

[4] T. Klingstrom and D. Plewczynski, "Protein-protein interaction and pathway databases, a graphical review," *Briefings in Bioinformatics*, vol. 12, no. 6, pp. 702–713, 2010.

[5] G. D. Bader, D. Betel, and C. W. Hogue, "BIND: the Biomolecular Interaction Network Database," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 248–50, 2003.

[6] S. Kerrien *et al.*, "The IntAct molecular interaction database in 2012," *Nucleic Acids Res.*, vol. 40, pp. D841–D8466, 2012.

[7] A. Chatr-aryamontri et al., "VirusMINT: a viral protein interaction database," *Nucleic Acids Res.*, vol. 37, pp. D669–D673, 2009.

[8] V. Navratil *et al.*, "VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks," *Nucleic Acids Res.*, vol. 37, pp. D661–D668, 2009.

[9] S. K. Kwofie, U. Schaefer, V. S. Sundararajan, V. B. Bajic, and A. Christoffels, "HCVpro: hepatitis C virus protein interaction database," *Infect. Genet. Evol.*, vol. 11, pp. 1971–1977, 2011.

[10] A. Chatr-Aryamontri *et al.*, "The BioGRID Interaction Database: 2013 update," *Nucleic Acids Res.*, vol. 41, pp. D816–D823, 2013.

[11] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The Database of Interacting Proteins: 2004 update," *Nucleic Acids Res.*, vol. 32, pp. D449–D451, 2004.

[12] T. S. K. Prasad *et al.*, "Human Protein Reference Database - 2009 Update," *Nucleic Acids Res.*, vol. 37, pp. D767–D772, 2009.

[13] D. Croft *et al.*, "Reactome: a database of reactions, pathways and biological processes," *Nucleic Acids Res.*, vol. 39, pp. D691–7, 2011.

[14] P. Shannon *et al.*, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–504, 2003.

[15] F. Ramírez, A. Schlicker, Y. Assenov, T. Lengauer, and M. Albrecht, "Computational analysis of human protein interaction networks," *Proteomics*, vol. 7, no. 15, pp. 2541–2552, 2007.

[16] A. Zhang, *Protein Interaction Networks: Computational Analysis*. Cambridge University Press, 2009, ch. 5.

[17] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science*, vol. 296, no. 5569, pp. 910–913, 2002.

[18] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 9, pp. 5116–5121, 2001.

[19] K. Kadota and K. Shimizu, "Evaluating methods for ranking differentially expressed genes applied to microArray quality control data," *BMC Bioinformatics*, vol. 12, no. 227, 2011.

[20] E. M. McCartney, K. J. Helbig, S. K. Narayana, N. S. Eyre, A. L. Aloia, and M. R. Beard, "Signal transducer and activator of transcription 3 (STAT3) is a pro-viral host factor for hepatitis C virus," *Hepatology*, to be published.

[21] M. L. Tornesello, L. Buonaguro, F. Tatangelo, G. Botti, F. Izzo, and F. M. Buonaguro, "Mutations in TP53, CTNNB1 and PIK3CA genes in hepatocellular carcinoma associated with hepatitis B and hepatitis C virus infections," *Genomics*, to be published.

[22] T. Simos, G. Thyfronitis, and C. Papaloukas, "Topological comparison of protein interaction networks between HBV and HCV," in *5th Panhellenic Conf. Biomed. Technol.*, paper 25, 2013.