Computational Biology and Bioinformatics
2017; 5(3): 36-42
http://www.sciencepublishinggroup.com/j/cbb
doi: 10.11648/j.cbb.20170503.12
ISSN: 2330-8265 (Print); ISSN: 2330-8281 (Online)

In-Silico Screening of Biomarker Genes of Hepatocellular Carcinoma Using R/Bioconductor

Afza Akbar, Mohd Murshad Ahmed, Safia Tazyeen, Aftab Alam, Anam Farooqui, Shahnawaz Ali, Md. Zubbair Malik, Romana Ishrat*

Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India

Email address:

romana05@gmail.com (R. Ishrat)

*Corresponding author

To cite this article:

Afza Akbar, Mohd Murshad Ahmed, Safia Tazyeen, Aftab Alam, Anam Farooqui, Shahnawaz Ali, Md. Zubbair Malik, Romana Ishrat. *In-Silico* Screening of Biomarker Genes of Hepatocellular Carcinoma Using R/Bioconductor. *Computational Biology and Bioinformatics.* Vol. 5, No. 3, 2017, pp. 36-42. doi: 10.11648/j.cbb.20170503.12

Received: July 26, 2017; **Accepted**: August 7, 2017; **Published**: August 25, 2017

Abstract: Hepatocellular Carcinoma is a primary malignancy of the liver. It is the fifth most common cancer around the world and is a leading cause of cancer related deaths. For about 40 years HCC has been predominantly linked with Hepatitis B and Hepatitis C infection. This work aims to find out potential biomarkers for HBV and HCV infected HCC through rigorous computational analyses. This was achieved by collecting gene expression microarray data from GEO (Gene Expression Omnibus) database as GSE series (GSE38941, GSE26495, GSE51489, GSE41804, GSE49954, GSE16593) and pre-processing it using Bioconductor repository for R. Following a robust mechanism including the use of statistical testing techniques and tools, the data was screened for DEGs (Differentially Expressed Genes). 3354 down regulated genes and 785 up regulated genes for HBV and 3462 down regulated and 251 up regulated genes for HCV were obtained. For a comparative study of DEGs from HBV and HCV, they were merged to look for potential biomarkers whose differential expression may result in carcinoma. A total of 17 biomarkers (1 up-regulated and 16 downregulated), was obtained which were further subjected to Cytoscape to generate a GRN using STRING app. Furthermore, module level analysis was performed as it offers robustness and a better understanding of complex GRNs. The work also focuses on the topological properties of the network. The results point out to the presence of a hierarchical framework in the network. They also shed a light on the interactions of biomarkers whose down regulation may result in HCC. These results can be used for future research and in exploring drug targets for this disease.

Keywords: HCC, HBV, HCV, DEGs, Hamiltonian Energy, Network Modelling

1. Introduction

Hepatocellular carcinoma, also known by the name of malignant hepatoma, is a primary malignancy of the liver. It represents a poor prognostic cancer and is the fifth most common cancer in the world [1]. The primary cause for this cancer appears to be chronic liver disease and liver cirrhosis [2]. Annually, the cancer is diagnosed in about more than half a million people worldwide. Early diagnose can sometimes be cured with surgery or transplant but in more advanced cases it cannot be cured. The few cases (less than 5%) of HCC that do not develop on the background of chronic liver disease, are diagnosed late and usually have poor chances of cure [3].

Age is an important factor for this cancer as the people of 50 years or more have a higher risk of HCC as compared to the young population. Interestingly, the rate of this malignancy is higher in males than in females. Being the fifth most common cancer in men it appears to be the seventh most common cancer in women [4]. For more than 30 years HCC has been predominantly associated with chronic infection of Hepatitis B and Hepatitis C virus. Thus, the problem is even more overwhelming in regions where the incidence of chronic viral hepatitis B and/or C is of high prevalence [5].

It should be noted that the occurrence of HCC is increasing in several developing nations and is likely to

increase in the same manner [6]. Although the rate of this tumor is low for the developed world, there is a distinct geographical variation in its incidence, with 81% of cases occurring in the developing world and 54% of these occurring in China [5]. In Chinese and black African population, mainly infected with HBV, the patients are younger, while in Sub-Saharan Africa (high incidence of HBV infection), where the incidence of HCC is the highest, it can appear in the third decade of life. In Asia and sub-Saharan Africa there are as many as 120 cases per 100,000. However for the Asia-Pacific region, it appears to be the third most common cause of cancer-related deaths [7].

According to the data from Surveillance Epidemiology and End Results (SEER), HCV infected HCC is a major cause of cancer mortality in the United States.

For a better understanding of the disease the biological data can be viewed using a computational approach and analysed accordingly. Comparative studies on HBV and HCV-infected HCC have shown that there exists distinct differential gene expression pattern (for each of them). In this work we use gene expression microarray data and analyse it to generate a Gene Regulatory Network using R and Cytoscape. We further find subnetworks and communities and then trace the biomarkers following the network. We also perform module enrichment and GO enrichment analysis.

2. Methodology

2.1. Pre-processing of Data

The microarray data was downloaded from GEO (Gene expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) database on NCBI. A stepwise search was performed for the identification of HCC-related gene expression profiles of humans using the keyword ‘hbv’ and ‘hcv’ for the corresponding GSE series. The datasets containing a comparison between normal and control tissues were preferred. A total of six datasets was downloaded (GSE38941, GSE26495, GSE51489; GSE41804, GSE49954, GSE16593) i.e. three for each type. For the pre-processing of microarray data Affy package was used. The packages were loaded onto the R environment from Bioconductor for data normalization and background correction was done [8]. Later, the text files generated were converted into a gene expression matrix with probe_id as rows and gene expression as the columns.

2.2. Screening for Differentially Expressed Genes

The gene expression matrix was used to obtain DEGs. The infected and normal controls were separated and average value of gene expression was calculated for each probe number. The probe numbers of the expression profile were later converted into the corresponding gene symbols following the correlation between gene and probe from the platform GPL570.

The fold change (FC) was calculated by subtracting the average values of infected samples from the corresponding values of normal controls [9]. A threshold of 0.5 was used for HBV and of 1 for HCV. Further, FC values were filtered to obtain DEGs. The common genes between HBV and HCV were found using VENNY 2.1.0. They were further screened for cancerous genes using NCG (Network of Cancer Genes, ncg.kcl.ac.uk) [10].

2.3. Network Construction and Topological Properties

The cancerous genes obtained from NCG (Network of Cancer Genes) were mapped onto Cytoscape to construct a gene regulatory network [11]. It is one of the common uses of Cytoscape to map attribute data onto a biological network, such as a protein-protein interaction network or metabolic pathway. The network was constructed using STRING database application (in the public database section) in Cytoscape [12]. Furthermore, the topological properties of the GRN were considered by constructing plots for degree distribution: node-degree distribution P(K), clustering coefficient C(K) and neighbourhood connectivity $C_N(K)$ and centralities: betweeness $C_B(K)$, closeness $C_C(K)$, eigen vector $C_E(K)$.

2.4. Finding Communities

The final and most important step of the process was to generate potential modules from the GRN so obtained. The modules were generated using R and Cytoscape simultaneously. An sif file was generated for the GRN from Cytoscape which was loaded onto R to be broken into communities. Text files containing the gene names of the respective modules were also generated through R by following a series of commands. The gene names were then used in Cytoscape to break the network into corresponding modules and biomarkers were traced following the hierarchy. This process was carried out until no communities were left that could be broken further and no genes were left to be traced. GO enrichment analysis was performed using DAVID 6.8 (Database for Annotation Visualization and Integrated Discovery, https://david.ncifcrf.gov/) [13].

3. Results and Discussion

3.1. Analysis of Topological Properties

The topological analysis of different large-scale biological networks highlights some recurrent properties: power law distribution of degree, scale-freeness, small world, which have been proposed to confer functional advantages such as robustness to environmental changes and tolerance to random mutations [14]. Network analysis acts as a powerful way for understanding the function and evolution of biological processes, provided that smaller functional modules are equally focused establishing a link between their topological properties and their dynamical behaviour.

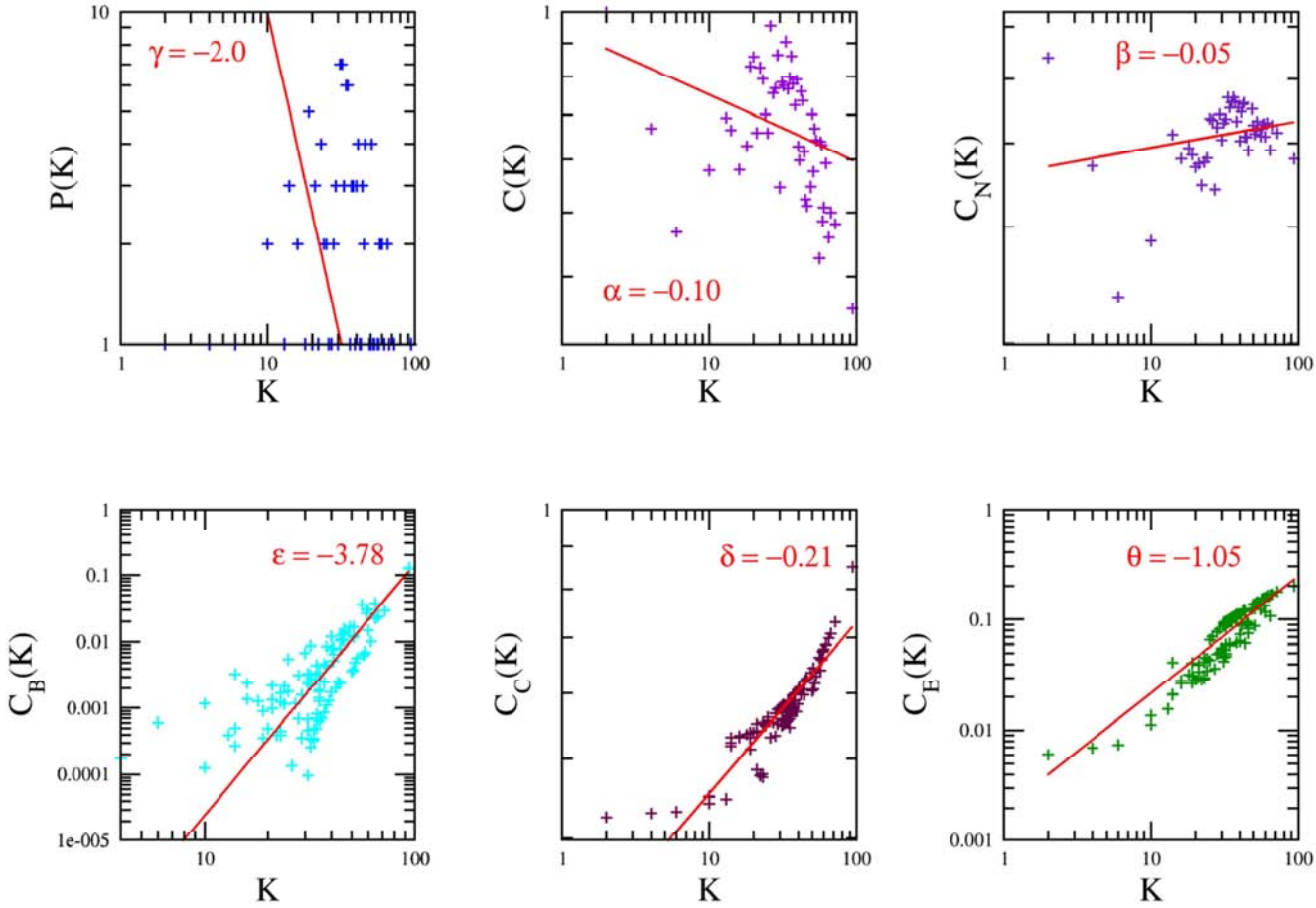

Figure 1. *Plots showing topological properties of GRN. First three plots representing degree distribution (P(K), C(K), $C_N(K)$) and last three plots representing centralities ($C_B(K)$, $C_C(K)$, $C_E(K)$).*

The topological parameters namely probability of degree distributions (P(K)), clustering co-efficient (C(K)) and neighbourhood connectivity ($C_N(K)$) exhibit power law and are analysed here.

The behaviour of the network is characterized by equations

(K) ~ $k^{-\alpha}$

(K) ~ $k^{-\beta}$

(K) ~ $k^{\emptyset}$

The +ve value in theta of connectivity parameter shows assortative nature of the network. While, the -ve value in alpha (α) of degree distribution shows availability of each node in the network. The -ve value in beta of clustering parameter shows dissociation in the communication between the nodes in network.

The basic centrality parameters, namely, betweeness $C_B(K)$, closeness $C_C(K)$, eigen vector $C_E(K)$ of the network also exhibit hierarchical behaviour given by,

$C_B(K)$ $\gamma = 3.78$

$C_C(K)$ $\delta = 0.21$

$C_E(K)$ $\mathcal{E} = 1.05$

The + ve values of the centralities exponents shows the strong regulating behaviour of the nodes in the network.

3.2. Gene Ontology (GO) Enrichment Analysis

To understand the DEGs so obtained, it is essential to have the knowledge of their specific function. The genes obtained from the microarray data were divided into up-regulated and down-regulated elements from which cancerous genes were found namely BUB1B, RUNX1T1, COL3A1, EGFR, FGFR2, GPC3, LAMA4, MKI67, NEK2, PEG3, PLCG2, R1T1, TTK, CCNB2, AKR1B10, CASC5 (up-regulated) and MALAT1 (down-regulated). For these genes, the Gene Ontology was found using DAVID and three important categories namely Biological Process (bp), Cellular Component (cc) and Molecular Function (mf) were noted. It gave the results as in Table 1.

Table 1. *Gene Ontology (GO) enrichment table showing bp, cc and mf.*

Symbol	GO_id	Biological Process (bp)	GO_id	Cellular Component (cc)	GO_id	Molecular Function (mf)
BUB1B	GO:0007091	metaphase/anaphase transition of mitotic cell cycle	GO:0000778	condensed nuclear chromosome kinetochore	GO:0004672	protein kinase activity

Symbol	GO_id	Biological Process (bp)	GO_id	Cellular Component (cc)	GO_id	Molecular Function (mf)
RUNX1T1	GO:0045892	negative regulation of transcription, DNA-templated	GO:0005634	Nucleus	GO:0003700	transcription factor activity, sequence-specific DNA binding
COL3A1	GO:0007160	cell-matrix adhesion	GO:0005578	proteinaceous extracellular matrix	GO:0005201	extracellular matrix structural constituent
EGFR	GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	GO:0016021	integral component of membrane	GO:0004714	transmembrane receptor protein tyrosine kinase activity
FGFR2	GO:0008284	positive regulation of cell proliferation	GO:0005654	Nucleoplasm	GO:0005007	fibroblast growth factor-activated receptor activity
GPC3	GO:0001658	branching involved in ureteric bud morphogenesis	GO:0005578	proteinaceous extracellular matrix	GO:0043395	heparan sulfate proteoglycan binding
LAMA4	GO:0001568	blood vessel development	GO:0005604	basement membrane	GO:0005102	receptor binding
MKI67	GO:0006259	DNA metabolic process	GO:0000775	chromosome, centromeric region	GO:0000166	nucleotide binding
NEK2	GO:0006468	protein phosphorylation	GO:0005813	Centrosome	GO:0004674	protein serine / threonine kinase activity
PEG3	GO:0000122	negative regulation of transcription from RNA polymerase II promoter	GO:0005654	Nucleoplasm	GO:0003676	nucleic acid binding
PLCG2	GO:0032237	activation of store-operated calcium channel activity	GO:0005886	plasma membrane	GO:0004871	signal transducer activity
RIT1	GO:0007265	Ras protein signal transduction	GO:0005622	Intracellular	GO:0005525	GTP binding
TTK	GO:0007093	mitotic cell cycle checkpoint	GO:0016020	Membrane	GO:0004712	protein serine / threonine / tyrosine kinase activity
CCNB2	GO:0000086	G2/M transition of mitotic cell cycle	GO:0005634	Nucleus	GO:0004693	cyclin-dependent protein serine/threonine kinase activity
AKR1B10	GO:0016488	farnesol catabolic process	GO:0070062	extracellular exosome	GO:0001758	retinal dehydrogenase activity
CASC5						

3.3. Network and Module Representation

A visual representation of the network and its corresponding sub-networks i.e. modules was done using Adobe Illustrator or AI. Each level was represented in different colours and the biomarkers in higher level of the network were highlighted in red. The result obtained marked the presence of hierarchy in the network. Table 2 shows the potential biomarkers that occurred in the higher levels of hierarchy. The hierarchical network is represented in Fig. 2.

Table 2. *Biomarkers highlighted in higher levels.*

Level	Module name	Gene name
Level 4	C1121	EGFR
Level 4	C1121	FGFR
Level 4	C1211	RUNX1T1
Level 4	C1211	AKR1B10
Level 4	C1212	MKI67
Level 4	C1212	PEG3
Level 4	C1213	RIT1
Level 4	C1221	GPC3
Level 4	C1222	COL3A1
Level 5	C1212a	PEG3
Level 5	C1212a	MKI67

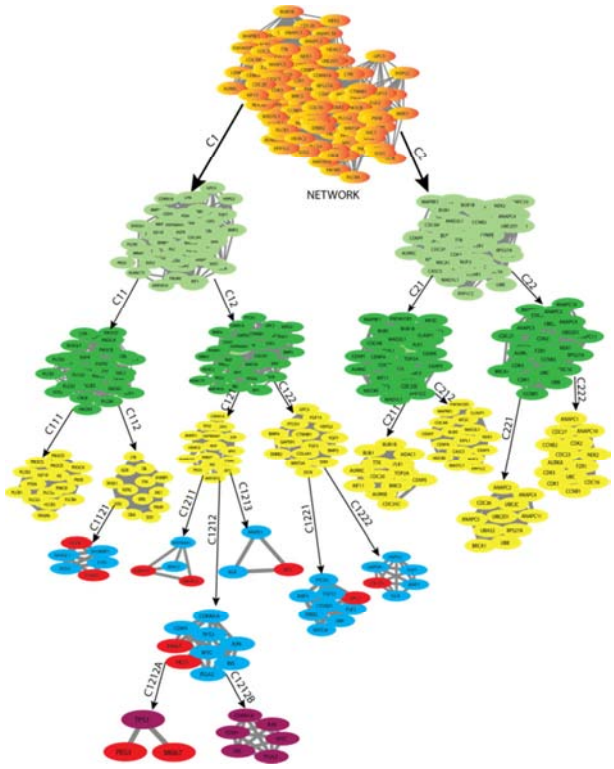

Figure 2. *Gene Network and its sub-networks or modules. The figure shows a five-level network where different colours are used for each level. Bio markers are highlighted in red at level 4 and 5.*

3.4. Network Properties and Gene Tracing

We can identify potential hubs from the network by knowing the maximum number of interactions each node has. The potential hubs identified in the network were: PEG3, MKI67, RUNX1T1, GPC3, FGFR2, EGFR, RIT1 and COL3A1. A plot for Hamiltonian energy was then generated for the potential biomarkers at different levels (Fig. 3.). The plot shows an active participation in the lower levels as compared to higher ones.

A plot was generated for modularity at different levels [15]. The plot shows a decreasing trend as it moves towards the higher levels, meaning modularity decreases from one level to the other and follows the same trend as we go on. Fig. 4. shows the plot for Modularity.

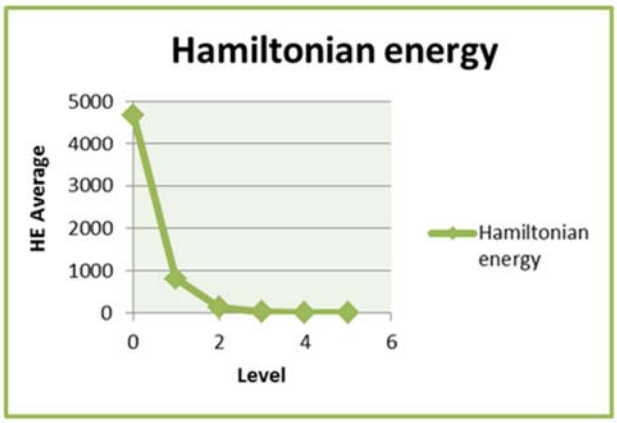

Figure 3. *Plot for Hamiltonian Energy of the system at different levels.*

Figure 4. *Modularity Plot showing a decreasing trend for modularity at consecutive levels.*

At first two communities emerge from the network i.e. C1 and C2 which further divide into two more communities. C1 has 10 biomarkers namely GPC3, COL3A1, EGFR, PLCG2, RIT1, PEG3, AKR1B10, RUNX1T1, MKI67, FGFR2 while C2 contains 5 biomarkers namely NEK2, TTK, CASC5, CCNB2, BUB1B. As the divergence increases each community divides into two except for C121, which diverges into 3 communities. The communities are represented in different colour for different levels (Fig. 5). EGFR, FGFR, RUNX1T1, RIT1 and AKR1B10 go till the fourth level. The community at fourth level i.e. C1212 diverges into 1212a and 1212b and gives rise to the last level that is the fifth level which has PEG3 and MKI67 in community 1212a.

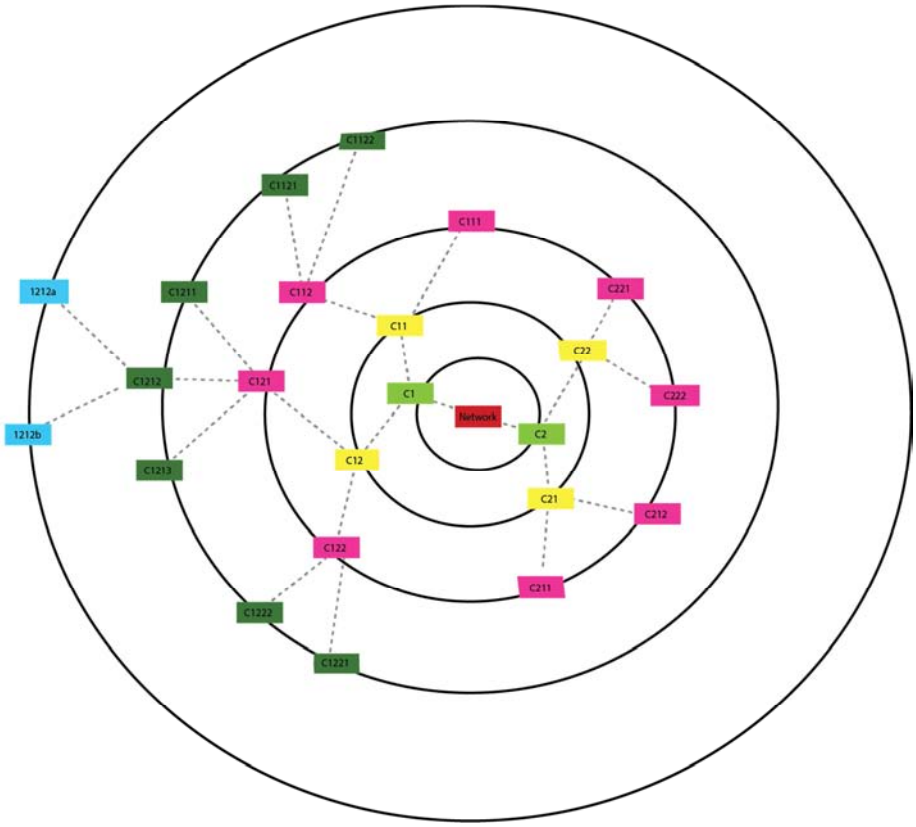

Figure 5. *Emergence of Modules at different levels of network.*

We then represent the hierarchy in the network and trace the potential biomarkers till the last level of the hierarchy. Fig. 6 represents the hierarchical network and the division of genes into different communities and modules. Each level is shown in a different colour for a better understanding.

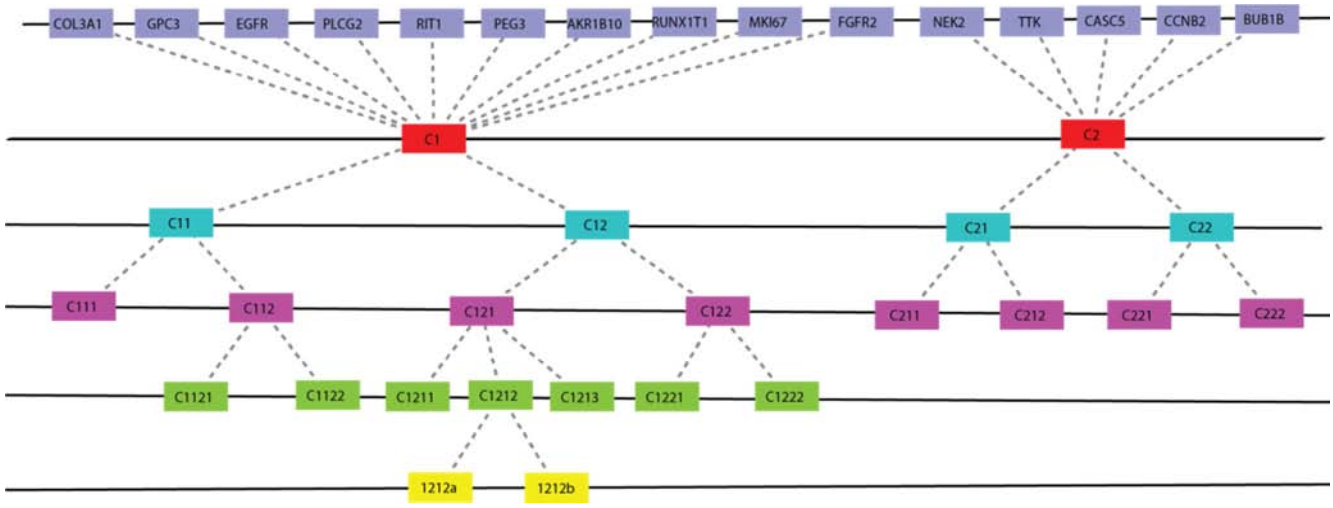

Figure 6. *Network hierarchy and Biomarkers dividing into different modules.*

4. Conclusion

This work aims at providing an insight into the regulatory network of HBV and HCV infected Hepatocellular Carcinoma. The communities (and sub-networks) provide an insight into the interactions of the biomarkers involved in causing the malignancy. It can be noted that potential biomarkers for HCC, like AKR1B10, COL3A1, FGFR2, EGFR, PEG3 and MKI67 are present till the higher levels of the network. Also the topological properties suggest two things: degree distribution plots (P(K), C(K), C_N(K)) suggest the presence of hierarchy in the network while the centralities (C_B(K), C_C(K), C_E(K)) suggest the assortative nature of the network. The presence of biomarkers in higher levels, the plot for modularity and the topological properties together suggest the presence of potential hubs at every level of the hierarchy of network. In conclusion, this work can be used for future research to explore for potential drug targets and when worked upon can provide methods for the development of better treatment against HBV and HCV infected HCC.

References

[1] K. J. Schmitz *et al.*, "Activation of the ERK and AKT signalling pathway predicts poor prognosis in hepatocellular carcinoma and ERK activation in cancer tissue is associated with hepatitis C virus infection," *J. Hepatol.*, vol. 48, no. 1, pp. 83–90, Jan. 2008.

[2] W. Yuan *et al.*, "Comparative analysis of viral protein interaction networks in Hepatitis B Virus and Hepatitis C Virus infected HCC," *Biochim. Biophys. Acta BBA - Proteins Proteomics*, vol. 1844, no. 1, pp. 271–279, Jan. 2014.

[3] H. B. El-Serag, "Epidemiology of Viral Hepatitis and Hepatocellular Carcinoma," *Gastroenterology*, vol. 142, no. 6, p. 1264–1273. e1, May 2012.

[4] V. W. Keng, D. A. Largaespada, and A. Villanueva, "Why men are at higher risk for hepatocellular carcinoma?," *J. Hepatol.*, vol. 57, no. 2, pp. 453–454, Aug. 2012.

[5] H. S. Te and D. M. Jensen, "Epidemiology of Hepatitis B and C Viruses: A Global Overview," *Clin. Liver Dis.*, vol. 14, no. 1, pp. 1–21, Feb. 2010.

[6] R. X. Zhu, W.-K. Seto, C.-L. Lai, and M.-F. Yuen, "Epidemiology of Hepatocellular Carcinoma in the Asia-Pacific Region," *Gut Liver*, vol. 10, no. 3, May 2016.

[7] S. A. Jones, D. N. Clark, F. Cao, J. E. Tavis, and J. Hu, "Comparative Analysis of Hepatitis B Virus Polymerase Sequences Required for Viral RNA Binding, RNA Packaging, and Protein Priming," *J. Virol.*, vol. 88, no. 3, pp. 1564–1572, Feb. 2014.

[8] A. Kauffmann, R. Gentleman, and W. Huber, "array Quality Metrics--a bio conductor package for quality assessment of microarray data," *Bioinformatics*, vol. 25, no. 3, pp. 415–416, Feb. 2009.

[9] M. A. Newton, "Detecting differential gene expression with a semiparametric hierarchical mixture method," *Biostatistics*, vol. 5, no. 2, pp. 155–176, Apr. 2004.

[10] S. Das and D. L. Mykles, "A Comparison of Resources for the Annotation of a *De Novo* Assembled Transcriptome in the Molting Gland (Y-Organ) of the Blackback Land Crab, *Gecarcinus lateralis*," *Integr. Comp. Biol.*, vol. 56, no. 6, pp. 1103–1112, Dec. 2016.

[11] M. D'Antonio, V. Pendino, S. Sinha, and F. D. Ciccarelli, "Network of Cancer Genes (NCG 3. 0): integration and analysis of genetic and network properties of cancer genes," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D978–D983, Jan. 2012.

[12] D. Szklarczyk *et al.*, "The STRING database in 2017: quality-controlled protein? protein association networks, made broadly accessible," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D362–D368, Jan. 2017.

[13] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000.

[14] S. Maslov, "Specificity and Stability in Topology of Protein Networks," *Science*, vol. 296, no. 5569, pp. 910–913, May 2002.

[15] J.-D. J. Han *et al.*, "Evidence for dynamically organized modularity in the yeast protein? protein interaction network," *Nature*, vol. 430, no. 6995, pp. 88–93, Jul. 2004.