The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010

Network Flow Model Based Analysis of Mouse Hepatocarcinogenesis

Yu-Qing Qiu[1,*] Xing-Ming Zhao[2] Zikai Wu[3]
Chaochao Wu[4] Rong Zeng[4,†] Luonan Chen[4,‡]

[1]Department of Chemical Pathology, the Chinese University of Hong Kong, Hong Kong, 999077, China
[2]Institute of Systems Biology, Shanghai University, Shanghai, 200444, China
[3]Business School, University of Shanghai for Science and Technology, Shanghai, 200093, China
[4]Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, 200031, China

Abstract Hepatocellular carcinoma (HCC) is one of the most harmful cancer in the world. The transgenic mouse model of hepatocarcinogenesis was established to decipher the pathological mechanism of HCC. Previous studies found that several transgenes play important roles in transgenic mouse carcinogenesis. However, the detail relationship between transgene and other genes is still unknown. In this paper, we propose a network flow based mathematical model to infer the transgene centering functional network (TCFN), which shows the regulatory effect of transgene to the differentially expressed genes, by integrating genome-wide high-throughput data and protein-protein interaction network. The proposed model was applied to transgenic mouse development study. In particular, tumor progress related TCFN on different development stages was identified. The analysis of dynamic changes of TCFN revealed some important genes highly correlated to the tumor development. The TCFN and the identified genes provide new insights on the mouse HCC pathogenic mechanisms.

Keywords Hepatocellular Carcinoma; Protein-Protein Interaction Network; Linear Programming

1 Introduction

Liver cancer is a major risk of human health, which causes 662,000 deaths worldwide per year, about half of them in China according to the World Health Organization (http://www.who.int/mediacentre/factsheets/fs297/en/). Since the liver is made up of many various cells, multiple types of tumors, which are malignant and benign, or metastatic tumor from different tissues, can develop in the liver. Hepatocellular carcinoma (HCC) is the major type of malignant primary liver neoplasm. It is the fifth most common cancer and the third most common cause of death from cancer worldwide [6].

*Email: qiuyuqingfred@cuhk.edu.hk.
†Corresponding author, email: zr@sibs.ac.cn.
‡Corresponding author, email: chen@eic.osaka-sandai.ac.jp.

The aetiological factors of HCC mainly include the virus infection (hepatitis B virus and hepatitis C virus), prolonged dietary aflatoxin exposure and alcoholic cirrhosis. In the molecular level, many genetic or epigenetic events associated with the development of HCC have been pointed out, such as the inactivate of the tumor suppressor P53, mutations in β-catenin, methylation of cancer-relevant genes (P16, COX2 *etc*.) and so on [5]. In addition, genomic instability (telomere shortening and chromosome segregation defects *etc*.) and genomic alterations (frequent chromosomal gains in 1q, and 6p, and losses in 1p and 4q, *etc*.) [5] are through to contribute to HCC. Although these mechanisms have been discovered, they are still far from diagramming the molecular, cellular and environmental mechanisms that drive disease pathogenesis.

To study the molecular pathogenical mechanism of HCC, a number of transgenic mice models of hapatocarcinogenesis have been established, in which selective expression of various cellular or viral genes in the liver induces a high predisposition to primary HCC development [3], *etc*. These mice invariably develop HCC with a relatively short period after birth. The transgene shows a peculiar expression pattern which greatly facilitates tumor onset.

Recently, along with the development of high-throughput mass spectrometry (MS) technologies, it extends the traditional method, such as western blot, from measuring a single protein to proteome-wide proteins simultaneously. This progress greatly facilitates and systematically investigate the behavior of all proteins, which is also suitable for the analysis of complex diseases, especially HCC. However, the interactions between proteins have not been considered with MS data. It is known that proteins function in an cooperative manner rather than isolated way. The proteins often interact together to affect the biological process or disease onset.

In this article, we systematically integrate the mass spectrometry data of protein expression and protein-protein interaction network to explore the heptocyte carcinogenesis of transgenic mice. Focusing on the important role of transgene and its related interaction network, we proposed an mathematical programming model to retrieve transgene centering functional interaction network (TCFN) at different development stage of transgenic mice. The topological structure of the identified network could uncover the functional linkage and cooperation between transgene and differentially expressed (DE) proteins. By analyzing the dynamical changes of TCFN, we got useful information to understand the process of HCC onset. Specifically, some topological important proteins were identified and confirmed to be related to cancer by literatures.

2 Materials

The in-house protein expression data of transgenic mouse liver tissue as well as age-matched normal mouse was measured by high-throughput proteomic mass spectrometry technique. In 6 time points i.e. 10 day after born, 2 month, 3 month, 5 month, 7 month and 11 month, which covered the entire process of hepatocarcinogenesis in transgenic mouse model, dynamical proteomic expression change of liver cell was detected respectively. At each time point, 6 mice composed of 3 transgenic mice and 3 normal mice were carefully chosen and subject to MS analysis. Finally, 3920 noredundant proteins expression profiles from 36 runs were obtained. For further analysis, the expression values were transformed by base 2 logarithm function.

The protein-protein interaction data used in this study is from STRING database (version 8) [8], which is a meta-resource that aggregates most of the available information on physical protein-protein interaction and functional linkage, such as co-expression, co-evolution etc. In this context, the protein-protein interactions are defined in the general sense which include all kind of direct functional relationship, such as physical or regulatory interaction. Each interaction is weighted by a scoring function which integrates information from numerous sources, including experimental repositories, computational prediction methods and text mining. To obtain more accurate results, interactions with weight $\geqslant 0.5$ were selected. There were totally 16566 proteins and 75314 interactions.

3 Methods

3.1 Differentially expressed proteins identification

To identify differentially expressed proteins between transgenic mice and wild-type mice in each time point, we use the fold-change (FC) measurement as follows:

$$\mathrm{FC}_i = \frac{\overline{x_i}}{\overline{y_i}},$$

where $\overline{x_i}$ and $\overline{y_i}$ represent the average expression value of protein i in transgenic samples and wild-type samples in each time point. Since sample size of both type of mice is 3 and there are many missing values in each samples, the statistical significance is hard to meet. Thus, comparing with other widely used methods in analysis of gene expression data, such as t-statistics, SAM [9] etc., FC is more suitable for this study. Due to high missing rate of raw data and plenty of minor expression changes, we set a threshold $\theta = 1.1$ to include more proteins. Although this might bring more random noise, the integration of protein-protein interaction network can partly eliminate such confounding effects. Proteins with FC score higher than θ are regarded as up-regulated, while proteins with FC score lower than $1/\theta$ are regarded as down-regulated. Up- and down-regulated proteins are all considered as differentially expressed (DE).

3.2 Functional networks construction

In each development stage of transgenic mice, DE proteins provide good signals to infer the molecular function changes. We constructed the network between transgene and DE proteins, called transgene centering functional network (TCFN), at each time point by integrating protein-protein interaction network and DE protein set. TCFN which contains transgene, DE proteins and direct or indirect interactions or pathways between them is a subnetwork of protein-protein interaction network. DE proteins are different when the time changes which lead to changes of topological structure of TCFN. As showed in results section, the dynamical behavior of TCFN would provide information undetectable by expression change.

To infer TCFN, we employ a flow model from graph theory [7, 1]. Formally, given an network $G = (V, E, W)$, where the set V of n nodes presents the proteins, the edge set E represents the interactions between proteins, and edge weight set $W = (w_{ij}), i, j = 1, \ldots, n$, represents the reliability of interactions. w_{ij} is an positive constant between 0 and 1, which lower w_{ij} indicates more reliability of interaction between protein i and j. For STRING interaction database used, the weight of interactions w'_{ij} provided in the database were

transformed by $w_{ij} = 1 - w'_{ij}$, and w_{ij} were set to infinity for protein i and j without interaction weight in STRING. In this edge-weighted network, we model the TCFN as a subnetwork connecting transgene with DE proteins with most reliable edges and find it by the following linear programming (LP) model:

$$\begin{aligned}
\min \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}w_{ij}, \\
\text{s.t.} \quad & \sum_{i=1}^{n} c_{si} = R, \\
& \sum_{j=1}^{n} c_{ji} - \sum_{k=1}^{n} c_{ik} = 0, \quad i = 1,\ldots,n, \quad \text{and} \quad i \notin \text{DEP}, \\
& \sum_{j=1}^{n} c_{ji} - \sum_{k=1}^{n} c_{ik} = 1, \quad i \in \text{DEP}, \\
& 0 \leqslant c_{ij} \leqslant R.
\end{aligned}$$

Node s is the source node which emits flows to the sink nodes which absorb flows. To uncover the activation force from transgene to DE protein (DEP), transgene is regarded as the source node and DEP mapped to the network G are regarded as sink nodes. Variable c_{ij} represents volume of flow from node i to node j, $R = |\text{DEP}|$ is the number of nodes in DEP set. The source node emit R units of flow to R sink nodes belonging to the DEP set, each of them absorbs 1 unit of flow. The subnetwork passed by flow connecting the source to the sinks is a functional network which connecting transgene to the DE proteins. The goal of this LP model is to find the most reliable TCFN, i.e. subnetwork with minimal weights. Thus, the object function is to minimize the weights of the subnetwork passed by flow. The first constraint presents that the source output R units of flow and the second constraint represents that every sink node consumes one unit of flow, while the third constraint is flow balance constraint which ensures every inter-medial node connecting the source and sinks without consuming any units of flow. The LP problem is solved by a free open source software lp_solv (http://lpsolve.sourceforge.net/). After obtaining the optimal solution, edges with $c_{ij} > 0$ are selected to construct the TCFN.

4 Results

Focusing on the transgene, by using the proposed LP model, we constructed transgene centering functional networks (TCFN) at each time point. In the middle stage of transgene silent, although transgene was overexpressed, it still performed biological functions and the TCFN was able to uncover its impact on DE proteins. As showed in Table 1, the size of TCFN including the number of nodes and edges were large in the early time (around 10 days, the number of nodes was 608 and the number of edges was 707). After this time, numbers were decreased to between 300 to 400 during 2 month and 5 month. In the last stage, from 7 months to 11 months, the network was increased to the one containing more than 1000 nodes. The size of the changes was correlated with the dynamic expression changes of transgene. The average clustering coefficient of nodes also had the same dynamical behavior of transgene. The power law test indicated that a TCFN was more likely to be scale free in the early stage and later stage, both the power law correlation and R^2 were higher than in the middle stage. This means that TCFN was scale-free [2] in

these two stage. It is well know that many complex real world network, such as Internet, social relation network, biological molecular interaction network etc. are scale-free. In such type of networks, the hubs, which were highest degree nodes, play important roles in network connectivity and robustness [2]. As showed below, hubs in the early and last stage might have more impact on tumor progression. However, we found that the average betweenness of nodes present an inverse change. In the middle period, betweenness was high, while it was low during early and last periods.

Table 1: Topological properties of TCFN

Time	# nodes	# edges	Avg. Clustering coefficient	Avg. Betweenness	Power law correlation	Power law R^2
10 days	608	707	0.027	0.075	0.728	0.801
2 months	352	360	0	0.106	0.655	0.722
3 months	360	360	0	0.109	0.691	0.744
5 months	471	497	0.003	0.1	0.656	0.707
7 months	632	705	0.015	0.087	0.748	0.794
11 months	1052	1469	0.038	0.075	0.779	0.862

In each TCFN, We analyzed hubs which are regard as important nodes. In complex networks, hubs are nodes with many neighbors, i.e. high degree. After counting the degree of each node, nodes with degree more than 10 in one time point was defined as hubs. Considering the degree of one node at each time point as a time serious, we clustered the hubs into clusters with similar dynamic properties by hierarchical clustering with correlation measure and single linkage strategy. As showed in Figure 1, there were 3 clusters observed. Large proportion of hubs were clustered with transgene into Cluster 2. The degree of them was high in the early and final time period, but low in the middle. While Cluster 1 containing only one gene Rb1 which has a particular profile that the degree during 10 days and 2 months is constantly low but high in 7 month and 11 month. The two time intervals might be its two different states. Rb1 is a tumor suppressor and regulator of cellular proliferation. Its loss of function often occurs relatively late in tumor progression [4]. Comparing with other tumor suppressing gene Trp53 which has similar profile with transgene, the lose of activation of Rb1 in the early stage may cause proliferation and tumor development. Beside, cluster 3 was diverse from other clusters that the degree was decreased after 10 days to an low level. These genes including Jun, Ephx1, Acat1, Gsmt4 and Gsmt1 might have important role in the neonatal mice to initiate carcinogenic process.

In the meantime, the betweenness [7] of each node was analyzed. Betweenness measures of the centrality of a node within a network that determines the relative importance of a node within the network. Nodes that occurred on many shortest paths between other nodes have higher betweenness than those that do not. We calculated all nodes' betweenness and selected nodes with betweenness higher than 0.1 at one time point for clustering analysis as above. The results are illustrated in Figure 2. Similar to hubs, many high betweenness nodes were clustered with transgene into one cluster. However, profiles diverse from transgene expression that in the middle period the betweenness is higher than early and last. One possible reason is that the proliferation signals from transgene are mainly transduced by them in the transgene inactive period. Since the proliferation does not stop this time, these genes, including Crk, Rb1, Jun, Trp53 and Mapk8, may cooperate

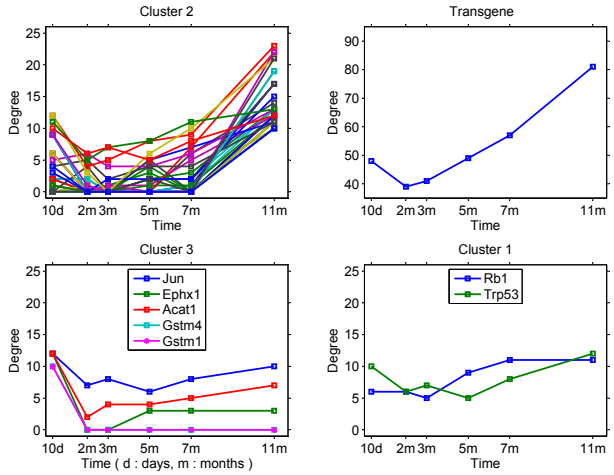

Figure 1: Clusters of hubs of similar degree change profile.

with transgene to elevate the tumor development. Another cluster containing Tnf, Gstm2, Gstp1, Cyp2c40 and ENSMUSP00000015983 had two peaks in 10 days to 2 months and 5 months to 7 months but low in 11 months. These genes might have two active period for carcinogenesis. In addition, Ptk2 composing one cluster had high activity in the early stage which might relate to tumor development initiation.

To analysis closely interacting nodes in the network, we calculated the clustering coefficient [7] of each node, which quantifies how close its neighbors are to being a clique (complete graph). In complex biological networks, nodes with high clustering coefficient always interact with their neighbors to form functional modules. By setting threshold of 0.1, we selected nodes with high clustering coefficient for clustering analysis. There were two mainly clusters which have inverse dynamical patterns (results not showed). One was active in the early period and the other was only active in the last stage. Functional annotation of these genes were related to metabolism and other molecular function, which was different from nodes with high degree or betweenness related to regulation and signal transduction. On the other hand, proteins with high clustering coefficient were more likely to form modules. The two clusters of proteins might function by forming modules, which are groups of proteins interacting closely to each other, to perform particular functions, in neonatal and tumor stage. It should be noted that the cluster of proteins active in the early stage, might be related to initiate the carcinogenic process.

5 Conclusion

Deciphering the molecular function and mechanism of genes or proteins in the HCC onset from the systematical perspective is an centering issue in the field of systems biology. To achieve this goal, approaches both in experiment and computation are important.

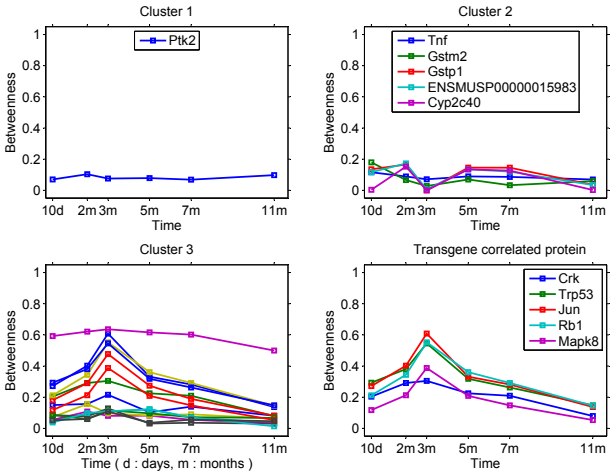

Figure 2: Clusters of high betweenness proteins.

The high-throughput technologies generate very large amount of data which need computational methods to integrate them, retrieval biological relevant information and then propose new hypotheses. In terms of this idea, this paper proposed a network based approach to identify transgene related functional networks related to carcinogenesis of transgenic mice. By combining the proteomic mass spectrometry data and interactomic data in a linear programming model, the regulatory effect of transgene to enhance tumor growth is inferred by the transgene centering functional networks. Comparing the TCFN in different stage, topologically important proteins which are not differentially expressed, such as cancer related gene Rb1 and Jun, are pointed out with important role. These results have provided useful insights for further functional mechanism validation. This indicates that the proposed approach is useful. As solving linear programming is not computationally costly, it is easy to apply to similar biological problems.

Acknowledges

This work is supported by the Chief Scientist Program of Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences with the grant number 2009CSP002. X.M. Zhao is partly supported by the Innovation Program of Shanghai Municipal Education Commission (10YZ01), Innovation Funding of Shanghai University and Shanghai Rising-Star Program (10QA1402700). The authors sincerely thank for reviewers' valuable comments.

References

[1] RK Ahuja, TL Magnanti, and JB Orlin. *Network flows: theory, algorithms, and applications*. Prentic Hall Inc., Englewood Cliffs (NJ), 1993.

[2] A.L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[3] F.V. Chisari, K. Klopchin, T. Moriyama, C. Pasquinelli, H.A. Dunsford, S. Sell, C.A. Pinkert, R.L. Brinster, and R.D. Palmiter. Molecular pathogenesis of hepatocellular carcinoma in hepatitis B virus transgenic mice. *Cell*, 59(6):1145–1156, 1989.

[4] J L Dean, A K McClendon, K R Stengel, and E S Knudsen. Modeling the effect of the RB tumor suppressor on disease progression: dependence on oncogene network and cellular context. *Oncogene*, 29(1):68–80, 2010.

[5] P.A. Farazi and R.A. DePinho. Hepatocellular carcinoma pathogenesis: from genes to environment. *Nature Reviews Cancer*, 6(9):674–687, 2006.

[6] Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, 33(suppl_1):D514–517, 2005.

[7] F. Harary. *Graph Theory*. Perseus, Cambridge, MA, 1995.

[8] L.J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, et al. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(Database issue):D412, 2009.

[9] G. V. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98(9):5116–5121, 2001.