Original Article

Dynamic protein-protein interaction subnetworks of lung cancer in cases with smoking history

Wei Yu¹, Li-Ran He², Yan-Chao Zhao¹, Man-Him Chan³, Meng Zhang¹ and Miao He¹

Abstract

Smoking is the primary cause of lung cancer and is linked to 85% of lung cancer cases. However, how lung cancer develops in patients with smoking history remains unclear. Systems approaches that combine human protein-protein interaction (PPI) networks and gene expression data are superior to traditional methods. We performed these systems to determine the role that smoking plays in lung cancer development and used the support vector machine (SVM) model to predict PPIs. By defining expression variance (EV), we found 520 dynamic proteins (EV>0.4) using data from the Human Protein Reference Database and Gene Expression Omnibus Database, and built 7 dynamic PPI subnetworks of lung cancer in patients with smoking history. We also determined the primary functions of each subnetwork: signal transduction, apoptosis, and cell migration and adhesion for subnetwork A; cell-sustained angiogenesis for subnetwork B; apoptosis for subnetwork C; and, finally, signal transduction and cell replication and proliferation for subnetworks D-G. The probability distribution of the degree of dynamic protein and static protein differed, clearly showing that the dynamic proteins were not the core proteins which widely connected with their neighbor proteins. There were high correlations among the dynamic proteins, suggesting that the dynamic proteins tend to form specific dynamic modules. We also found that the dynamic proteins were only correlated with the expression of selected proteins but not all neighbor proteins when cancer occurred.

Key words Lung cancer, protein-protein interaction (PPI), network

Lung cancer is currently the deadliest disease in the world. Smoking is the primary cause of lung cancer and has been linked to 85% of cases. The occurrence and development of this disease is a multi-gene, multi-stage, and extremely complex process that involves several changes, including oncogene activation, tumor suppressor gene mutation and deletion, tumor cell apoptosis suppression, and microsatellite instability^[1-3].

Generally, many factors within the tumor microen-

vironment can influence cellular metabolism, signal transduction, and gene expression. However, most studies on lung cancer pathogenesis focus primarily on a single or limited number of genes and simple functional annotation using standard research methods. Only at the system level can the molecular mechanisms of cancer be revealed effectively. In the protein-protein interaction (PPI) network, the dynamic modules or subnetworks of proteins may have leading roles in the cancer development and metastasis process. The static modules of proteins may belong to the inherent components in a PPI network; these modules tend to associate with the “noises” of protein expression, genetic modification, and genetic evolution. The static modules of proteins may be a buffer in the variation of the PPI network, and cells having these proteins are robust^[4]. Thus, it is very important to explore the dynamic PPI subnetwork of lung cancer in patients with smoking history at the cellular level.

Authors' Affiliations: ¹School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong 510275, P. R. China; ²College of Software Technology, South China Agricultural University, Guangzhou, Guangdong 510642, P. R. China; ³Department of Architecture, Cardiff University, Cardiff, Wales, CF10, the United Kingdom.

Corresponding Author: Miao He, School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong 510275, P. R. China. Tel: +86-20-84110036; Email: lsshem@mail.sysu.edu.cn.

doi: 10.5732/cjc.012.10099

Data and Methods

Data sources

All human protein sequence data (9289 in total) and interaction data (37 066 pairs in total) were downloaded from the Human Protein Reference Database (HPRD)^[5], Release 8, 1 Sep 2009.

Gene expression data for lung cancer patients with smoking history and non-cancer samples were downloaded from the Gene Expression Omnibus (GEO) Database of the National Center for Biotechnology Information (NCBI)^[6]. The original data, GSE94115, were obtained from the smoking groups^[7]. A total of 72 patients with lung cancer were randomly selected, and 72 individuals without lung cancer were randomly selected as the training set. Another 17 patients with lung cancer and 17 individuals without lung cancer were randomly selected as the test set (Table 1).

Protein expression variance

Expression variance (EV) can be used to measure the dynamic expression of genes in the genome. The value of EV is the percentage of gene expression variance divided by the genome expression variance. If EV is small, the difference between differential expression of two genes in the genome is also small^[7]. Thus, by defining the EV value, we could also classify proteins coding by the genes as dynamic or static. Here, if $EV > 0.4$, the protein was classified as dynamic, whereas if $EV < 0.2$, the protein was classified as static.

Pearson correlation

The correlation of expression between two proteins was measured by determining the Pearson correlation coefficient (PCC). The smaller the absolute value of PCC, the lower the correlation of expression. We referred to the screening criteria of Goh *et al.*^[8] to define the values. If the absolute value of PPC was greater than 0.6 and the EV value was simultaneously greater than 0.4, the related proteins were selected to compose

the dynamic PPI subnetwork.

PCC was calculated with the following formula:

$$PCC = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

Here, the vector (x_i, y_i) represents the interaction pair of proteins A and B, respectively; x and y and represent the average expression of proteins A and B, respectively, in the 72 samples.

PPI prediction and subnetwork visualization

The support vector machine (SVM) model was used to predict PPIs. Subnetworks were visualized using Cytoscape software with the Cerebral plug-in, which could locate proteins interacting in cells^[9].

Results

Dynamic PPI subnetworks

Based on the data in Table 1, we calculated the parameters (C, g) of the SVM model to predict PPIs. (C, g) equaled (2, 0.03125), and the 5-fold calibration accuracy validation rate was 79%. The forecast accuracy rate was 70.58%, though only when test data were used.

In total, we identified 520 dynamic proteins ($EV > 0.4$) and 2754 static proteins ($EV < 0.2$), and we successfully built dynamic PPI subnetworks of lung cancer in patients with smoking history (Figure 1).

Functions of the dynamic PPI subnetworks

Using the Gene Ontology database, we determined that the majority of proteins in the subnetworks functioned in the physiological processes of cell migration and adhesion, apoptosis, signal transduction, cell-sustained angiogenesis, and cell replication and proliferation. The primary collective functions of each PPI subnetwork could be stamped by the key nodes of

Table 1. The original data and classes from the Gene Expression Omnibus (GEO) Database that used to predict protein-protein interactions by support vector machine (SVM) model

Serial number	Unit	Training/testing	Cancer/non-cancer
GSM94020-GSM94075, GSM94155-GSM94172	72	Training	Cancer
GSM94767-GSM94784	17	Testing	Cancer
GSM940100-GSM940148 GSM94077-GSM94099	72	Training	Non-cancer
GSM94785-GSM94801	17	Testing	Non-cancer

Figure 1. Dynamic protein-protein interaction subnetworks A-G. A, the functions of subnetwork A are mainly signal transduction, apoptosis, and cell migration and adhesion; B, the function of subnetwork B is mainly cell-sustained angiogenesis; C, the primary function of subnetwork C is apoptosis; D-G, the functions of subnetworks D-G are mainly signal transduction, cell replication and proliferation.

proteins or a group of proteins with the similar functions in the specific physiological processes in PPI subnetwork.

Analyzing the PPIs and pathways shown in Figure 1A, we determined that the functions of subnetwork A were mainly signal transduction, apoptosis, and cell migration and adhesion (Table 2). This suggests that the pathways in subnetwork A played central role in cancer cell-to-cell communication, cancer cell apoptosis control, and cancer cell adhesion and invasion to other normal tissues.

The function of subnetwork B (Figure 1B) was mainly cell-sustained angiogenesis (Table 2). As we know, cancer cells require nutrients to grow, and these nutrients travel to cancer cells through blood vessels. Therefore, blood vessel and tissue generation are important factors of cancer development. Subnetwork B showed some key pathways of blood vessel generation.

The primary function of subnetwork C was apoptosis (Table 2, Figure 1C). Proteins in this subnetwork such as fragile X mental retardation related protein 1 (FXR1), dynein light chain LC8-type 1 (DYNLL1), and heat shock

10-kDa protein 1 (HSPE1) positively regulated programmed cell death, whereas annexin A5 (ANXA5) and hepatitis B virus x interacting protein (HBXIP) negatively regulated programmed cell death^[10].

Subnetworks D–G (Figures 1D–G) functioned mainly in signal transduction, cell replication, and proliferation (Table 2). Generally, tumor cells generate many of their own growth signals, thereby reducing their dependence on stimulation from their normal tissue microenvironment. While most soluble mitogenic growth factors (GFs) are made by one cell type to stimulate proliferation of another, many of the growth signals can drive the proliferation of carcinoma cells, thus allows cell proliferation in specific phase of tumor development^[11]. Subnetworks D–G presented partial pathway of tumor proliferation.

Evaluation of the dynamic proteins

Because whole human PPI networks are incomplete for lung cancer patients with smoking history, we could not clearly determine the effects of each lung cancer-related protein and pathway. Although we built several dynamic PPI subnetworks, we still did not know

the role that these subnetworks play in processes of tumor development. Thus, to evaluate the effects of dynamic proteins, we measured the probability distributions of the degree of the dynamic protein and the static protein respectively. We found that the probability distributions between the degree of the dynamic protein (Figure 2A) and the static protein (Figure 2B) were very different and that the maximum density of the degree of the dynamic protein was relative lower, suggesting that the dynamic proteins were not the core proteins which widely connected with their neighbor proteins.

We found that the probability distributions between the average EV of neighbor proteins of the dynamic proteins (Figure 2C) and the static proteins (Figure 2D) had no significant difference. But there existed difference between the probability distributions of the average PCC of neighbor proteins in the dynamic proteins (Figure 2E) and static proteins (Figure 2F). For the average EV of neighbor proteins of the dynamic proteins, the maximum density of the probability distribution was low, suggesting that the expression difference between two neighbor proteins was also small. For the average PCC of neighbor proteins of the dynamic proteins, the mean value of the probability distributions was large,

Table 2. Functional classification of proteins in protein-protein interaction subnetworks A–G

Subnetwork	Function	Proteins
A	Cell-sustained angiogenesis	LMO2
	Cell adhesion and migration	RAF1, TJP1, TJP2, CTNNA1, CSDA, PPP1CC
	Apoptosis	HSPA1B, HSPA1A, RAF1, STAT1, YWHAZ, BAG3, YWHAQ
	Signal transduction	NMI, STAT1, IFNGR1, IRF2, ISGF3G, RAF1, HSPA1B, RHEB, SHOC2
	Cell replication and proliferation	RAF1
B	Cell-sustained angiogenesis	PAFAH1B1, HIF1A, RAB1A, GOLGA5
	Cell adhesion and migration	PAFAH1B1, PGL1, HIF1A
	Apoptosis	MIF
	Signal transduction	HIF1A
	Cell replication and proliferation	PGK1, MIF, CAPNS
C	Cell-sustained angiogenesis	ACTG1, DYNLL1
	Cell adhesion and migration	ACTG1, PAPBPC4
	Apoptosis	FXR1, DYNLL1, ANXA5, HSPE1, HBXIP
	Signal transduction	HSPD1, APLP2
	Cell replication and proliferation	PPIA, HSPD1
D	Signal transduction	PDCD6IP
	Cell replication and proliferation	PDCD6
E	Cell adhesion and migration	TUBB
	Apoptosis	TUBB
	Cell replication and proliferation	FTH1
F	Signal transduction	PRKAR1A, AKAP11
G	Cell adhesion and migration	IQGAP1
	Signal transduction	CALM1, RRAD
	Cell replication and proliferation	DDX5

Figure 2. Dynamic and static proteins' probability distributions of three main parameters—the degree of proteins, the average EV of neighbor proteins, and the average PCC of neighbor proteins. A, C, E, dynamic proteins; B, D, F, static proteins.

suggesting that there were high correlations among the dynamic proteins. The results showed that three main parameters—the degree of proteins, the average EV of neighbor proteins, and the average PCC of neighbor proteins—could basically reflect the relationship between proteins and surrounding proteins and were related to changes in protein expression. Moreover, the dynamic proteins, as well as static proteins, might have a similar correlation with surrounding proteins. These results suggest that the dynamic proteins were only correlated with the expression of selected proteins but not all neighbor proteins when cancer occurred.

We retrieved the functional annotations of selected dynamic proteins from the Gene Ontology Database and assessed their functional link with lung cancer pathogenesis (Table 3). Each of these genes and their protein products are related to cancer pathogenesis, though not specifically to lung cancer.

Discussion

In this study, we found 520 dynamic proteins and 2754 static proteins using the data from HPRD and the GEO database. We also built 7 dynamic PPI subnetworks of lung cancer in patients with smoking history. Initial analysis revealed the main functions of each PPI subnetwork: signal transduction, apoptosis, and cell migration and adhesion for subnetwork A; cell-sustained angiogenesis for subnetwork B; apoptosis for subnetwork C; and signal transduction and cell replication and proliferation for subnetworks D–G. These subnetworks

reveal potential mechanisms underlying lung cancer development^[12]. For the main parameter—the degree of proteins, the probability distribution of dynamic proteins and static proteins was different, obviously showing that dynamic proteins were not the core proteins which widely connected with their neighbor proteins. There were high correlations among the dynamic proteins, suggesting that the dynamic proteins tend to form specific dynamic modules.

Systems approaches that combine human PPI networks and gene expression data are superior to traditional methods, which can only analyze small amounts of gene expression data. Here, we compared high-throughput microarray expression data of 72 healthy smokers and 72 smokers with lung cancer, and we built several human dynamic PPI subnetworks. The gene expression data were then mapped on dynamic PPI subnetworks.

We calculated each protein's EV value according to gene expression changes in lung cancer patients with smoking history and healthy samples. EV value was used to define the dynamic protein with biggest expression change or the static protein with smallest expression change. Based on the relationship between protein expression and the PPI network, we analyzed the functions of the dynamic PPI subnetworks. By analyzing the degree of relation between dynamic proteins, static proteins, and their neighbor proteins, as well as the average EV and average PCC of neighbor proteins, we were able to evaluate the effects of the dynamic PPI subnetworks in cancer development.

The dynamic proteins that we identified represented

Table 3. Functional annotations of other partial dynamic proteins retrieved from Gene Ontology (GO) Database

Function from GO annotation	Proteins
Signal transduction	TOLLIP, HINT1, MAPK11, TNFAIP3, IL22, SOCS5, CORO2A, PDZD3, EEF1E1, RANBP2, PDPK1, MAPK6, PTEN, MPP3
Ion/glucose/transmembrane/vesicle-mediated/intracellular transport	NUDT9, SLC5A1, NDUFV2, MYO5A, COX7A2L, SRP54, PDZD3, ARCN1, SLC25A11, CPNE3
Response to stimulus	TOLLIP, IL22, EIF2B3, PDZD3, PEF1, IL1R2
Regulation of apoptosis	EEF1E1, PTEN, DAD1, NDUFS1, TNFAIP3
Regulation of transcription via RNA polymerase II promoter	ORC2L, ECD, SOX9, KLF9, EGR1, HSBP1, SAP30
Ubiquitin-dependent protein catabolic process	CDC16, PSMD6, PSMD10, TSG101, UCHL3
Tissue/organism development	DDX1, KRT85, NDUFV2
DNA replication and damage response	ORC2L, ORC5L, EEF1E1
Regulation of cell proliferation	CDC16, EEF1E1, PTEN
Regulation of cell cycle process	CDC16, PSMD6
Establishment of vesicle location/Golgi transport vesicle coating	TMED10, COPB2, ARCN
Macromolecular complex assembly	EIF2B3, EPRS, DDX1
Cell adhesion and motion	PTEN
Cellular respiration and homeostasis	NDUFS1

all essential functions of lung cancer development, which included maintenance of intracellular dynamic balance and regulation of programmed cell death, cell movement and localization, cell proliferation, immunoreactions, and transcription initiation via RNA polymerase II. Particularly, these essential functions were related to chemical or physical injury-induced inflammatory reactions and chemical stimulus reactions, which suggests that the cellular damage caused by smoking was the critical factor leading to lung cancer^[13,14]. In other words, lung cancer in patients with smoking history may be caused by proteins with a high EV value that function in the transition from precancerous stage to metastatic stage.

Because the dynamic proteins linked to lung cancer did not show different degrees, and because we observed different average EV and PCC values of neighbor proteins from the static protein, we found that not all the dynamic proteins were at core positions of the PPI networks or were not Hub nodes. Our finding that dynamic proteins did not show higher tendency of distribution than static proteins was in keeping with the previous conclusions of Goh *et al.*^[8], which indicates that vast majority of disease genes were nonessential and

showed no tendency to encode Hub proteins. Our discovery verified the feasibility of seeking cancer pathways.

However, the proteins that we used to build the dynamic PPI subnetworks were only a part of the complete human PPI network, thus the accuracy of PPI prediction with the SVM model needs to be further improved.

In conclusion, by analyzing dynamic protein subnetworks, we found that the proteins of such subnetwork took part in cancer-related biological processes. These proteins related to lung cancer could be used to predict cancer-related pathways.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (No. 91130009) and Science and Technology Planning Project of Guangdong Province of China (No. 2003A3080503).

Received: 2012-04-18; revised: 2012-09-07;
accepted: 2012-09-21.

References

- [1] Plebani M, Basso D, Navaglia F, et al. Clinical evaluation of seven tumour markers in lung cancer diagnosis: can any combination improve the results? *Br J Cancer*, 1995,72:170–173.
- [2] Vielhaber S, Jakubiczka S, Gaul C, et al. Brain 1H magnetic resonance spectroscopic differences in myotonic dystrophy type 2 and type 1. *Muscle Nerve*, 2006,34:145–152.
- [3] Beane J, Sebastiani P, Liu G, et al. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol*, 2007,8:R201.
- [4] Tang X, Wang J, Liu B, et al. A comparison of the functional modules identified from time course and static PPI network data. *BMC Bioinformatics*, 2011,12:339.
- [5] Keshava Prasad TS, Goel R, Kandasamy K, et al. Human protein reference database—2009 update. *Nucl Acids Res*, 2009,37:D767–D772.
- [6] Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: archive for functional genomics data sets—10 years on. *Nucl Acids Res*, 2010,39:D1005–D1010.
- [7] Komurov K, White M. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Mol Syst Biol*, 2007,3:110.
- [8] Goh KI, Cusick ME, Valle D, et al. The human disease network. *Proc Natl Acad Sci*, 2007,104:8685–8690.
- [9] Barsky A, Gardy JL, Hancock REW, et al. Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics*, 2007,23:1040–1042.
- [10] Marusawa H, Matsuzawa S, Welsh K, et al. HBXIP functions as a cofactor of survivin in apoptosis suppression. *Embo J*, 2003, 22:2729–2740.
- [11] Hanahan D. The hallmarks of cancer. *Cell*, 2000,100:57–70.
- [12] Zhou TT. Network systems biology for targeted cancer therapies. *Chin J Cancer*, 2012,31:134–141.
- [13] Yildirim MA, Goh KI, Cusick ME, et al. Drug-target network. *Nat Biotech*, 2007,25:1119–1126.
- [14] Lee G, Walser TC, Dubinett SM. Chronic inflammation, chronic obstructive pulmonary disease, and lung cancer. *Curr Opin Pulm Med*, 2009,15:303–307.