Chinese Journal of Cancer

Original Article

Using the theory of coevolution to predict protein-protein interactions in non-small cell lung cancer

Meng Zhang[1], Man-Him Chan[2], Wen-Jian Tu[1], Li-Ran He[3], Chak-Man Lee[4] and Miao He[1]

Abstract

Systems biology has become an effective approach for understanding the molecular mechanisms underlying the development of lung cancer. In this study, sequences of 100 non-small cell lung cancer (NSCLC)–related proteins were downloaded from the National Center for Biotechnology Information (NCBI) databases. The Theory of Coevolution was then used to build a protein-protein interaction (PPI) network of NSCLC. Adopting the reverse thinking approach, we analyzed the NSCLC proteins one at a time. Fifteen key proteins were identified and categorized into a special protein family F(K), which included Cyclin D1 (CCND1), E-cadherin (CDH1), Cyclin-dependent kinase inhibitor 2A (CDKN2A), chemokine (C-X-C motif) ligand 12 (CXCL12), epidermal growth factor (EGF), epidermal growth factor receptor (EGFR), TNF receptor superfamily, member 6 (FAS), FK506 binding protein 12-rapamycin associated protein 1 (FRAP1), O-6-methylguanine-DNA methyltransferase (MGMT), parkinson protein 2, E3 ubiquitin protein ligase (PARK2), phosphatase and tensin homolog (PTEN), calcium channel voltage-dependent alpha 2/delta subunit 2 (CACNA2D2), tubulin beta class I (TUBB), SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin, subfamily a, member 2 (SMARCA2), and wingless-type MMTV integration site family, member 7A (WNT7A). Seven key nodes of the sub-network were identified, which included PARK2, WNT7A, SMARCA2, FRAP1, CDKN2A, CCND1, and EGFR. The PPI predictions of EGFR-EGF, PARK2-FAS, PTEN-FAS, and CACNA2D2-CDH1 were confirmed experimentally by retrieving the Biological General Repository for Interaction Datasets (BioGRID) and PubMed databases. We proposed that the 7 proteins could serve as potential diagnostic molecular markers for NSCLC. In accordance with the developmental mode of lung cancer established by Sekine *et al*., we assumed that the occurrence and development of lung cancer were linked not only to gene loss in the 3p region (WNT7A, 3p25) and genetic mutations in the 9p region but also to similar events in the regions of 1p36.2 (FRAP1), 6q25.2–q27 (PARK2), and 11q13 (CCND1). Lastly, the invasion or metastasis of lung cancer happened.

Key words Coevolution, non-small cell lung cancer (NSCLC), protein-protein interactions (PPI)

Primary lung cancer can be categorized into small cell lung cancer (SCLC) or non-small cell lung cancer (NSCLC). The most common types of NSCLC are squamous cell carcinoma, large cell carcinoma, and adenocarcinoma. In the clinic, approximately 20% of lung cancer patients are diagnosed with SCLC, and approximately 80% are diagnosed with NSCLC.

The biological behavior of SCLC and NSCLC is significantly different. These two histopathologically distinct types of lung cancer grow and disseminate in different ways and are treated differently. SCLC is highly malignant and is characterized by rapid proliferation and metastasis. Some NSCLC tumors grow and spread more

Authors' Affiliations: [1]School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong 510275, P. R. China; [2]Department of Architecture, Cardiff University, Cardiff, Wales, the United Kingdom; [3]College of Software Technology, South China Agricultural University, Guangzhou, Guangdong 510642, P. R. China; [4]BEEXergy Consultant Ltd, Hong Kong, P. R. China.

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.012.10100

slowly, making them le ss prone to developing early metastases and more amenable to surgical treatment during the early stages of the disease[1].

Due to advancing experimental techniques and the application of high-throughput methods in the post-genomic era, it was possible to research protein-protein interactions (PPI) and the networks to which the proteins belong. Systems biology is currently the most effective approach for understanding the molecular mechanisms of lung cancer. The use of bioinformatics, combined with proteomics approaches, enables the identification of unknown protein functions as well as new functions for familiar proteins based on the PPI analysis. Identifying the key nodes of proteins would be helpful in revealing the molecular mechanisms underlying lung cancer. The development, evaluation, and application of the protein-based diagnostic approach would enable predicting patient susceptibility to lung cancer and identifying diagnostic molecular markers to detect early stages of lung cancer. Furthermore, this approach would be able to predict the response of lung tumors to drugs. It is also of great importance to develop calculation-based approaches that use molecular simulations to aid drug design, which would shorten the research time required for the development of new drugs and support the generation of effective and novel treatment approaches for lung cancer, thereby improving the health of all people.

Recently, a PPI network of Nanog functions was constructed, and the functional relevance of some newly identified Nanog components were validated[2]. Another PPI network has been developed for 54 proteins derived from 23 inherited ataxias. The network was expanded by incorporating literature-curated and evolutionarily conserved interactions, and 770 predominantly novel PPIs were identified using a stringent yeast two-hybrid screen. Many ataxia-causing proteins shared interaction partners, a subset of which has been implicated in neurodegeneration using animal models[3].

The cancer structural protein interface network (ciSPIN), a cancer interaction network, was generated by integrating protein interfaces. The results revealed that cancer-related proteins have smaller, more planar, more charged, and less hydrophobic binding sites than non-cancer proteins, indicating that cancer-related proteins are characterized by low affinity and high specificity interactions. The affinity of interactions between the proteins of the multi-interface hub tended to be higher than that between the proteins of the single-interface hub[4]. The Center for Cancer Systems Biology Human Interactome Version 1 (CCSB-HI1) dataset, a new version of the proteome-scale map of human binary PPIs, revealed more than 300 new connections with over 100 disease-associated proteins[5].

In this paper, we focused on NSCLC-related proteins. The coevolution approach was adopted to analyze NSCLC PPIs, and PPI networks were developed. We hope to explore the molecular mechanisms of NSCLC and to provide a guideline for future searches of the key proteins and preliminary diagnostic molecular markers for NSCLC.

Data and Methods

Data sources

Sequences of 100 genes and 100 related coding proteins related to NSCLC were downloaded from the National Center for Biotechnology Information (NCBI) databases (http://www.ncbi.nlm.nih.gov/).

Coevolution Theory

The coevolution theory is an approach used to study the PPIs that are based on the coevolutionary relationships of proteins. Coevolution describes the evolutionary choice that enables similar proteins to either co-exist or not co-exist within multiple genomes. If several proteins exist in a macromolecular component, or participate in a metabolic or signal transduction process, the relationships between these proteins are known as "functional linkages".

The premise of coevolution is that proteins that have similar functions and effects also have similar phylogenetic trees. The sequences of lung cancer-related proteins are compared with homologous sequences from other species, and the distance matrix between these proteins is calculated using Bioedit, a free software released by Tom Hall. The Evolution Theory speculates that two queried proteins may have a common evolutionary ancestor and serve as the basis for sequence alignments, which determine whether there exists sufficient similarity between two sequences. Pairwise sequence alignment determines whether there exists a molecular evolutionary relationship between two sequences by comparing similar regions and conserved sites. Then, based upon the alignment, the distance matrix determines the distance relationships between the two proteins. The distance can be calculated by equation

$$d(i,j) = 1 - \frac{S(i,j) - S_r(i,j)}{S_{\max}(i,j) - S_r(i,j)} \quad (1),$$

where d (i, j) represents the distance between sequence i and sequence j; S (i, j) represents the weighted sum of the scores for each alignment position of sequences i and j; S_r (i, j) represents the weighted sum of the alignment score after the randomization of sequences i and j; and S_{max} (i, j) represents the maximum

possible alignment score between two protein sequences (when the two sequences are the same, the maximum is used). The normalized distance between two sequences is represented by a calculated numerical value between 0 and 1; if the two sequences are identical, the distance will be 0. If the two sequences are greatly different, the distance value nears 1[6].

Upon calculating the distances between lung cancer–related proteins and the homologous proteins of other species, including *Bos Taurus*, *Rattus norvegicus*, *Canis familaris*, *Mus musculus* and *Danio rio*, the corresponding evolutionary distance matrix can be generated and used to predict the possibility of the interaction between two proteins. Equation (2) is used to calculate the correlation coefficients (r) between the evolutionary distance matrixes of the proteins,

$$r = \frac{\sum_{i=1}^{n} \left(R_i - \overline{R}\right)\left(S_i - \overline{S}\right)}{\sqrt{\sum_{i=1}^{n} (R_i - \overline{R})^2}\sqrt{\sum_{i=1}^{n} (S_i - \overline{S})^2}} \quad (2),$$

where r is the linear correlation coefficient; n=(N^2-N)/2, where n is the number of sequences entered into the multiple sequence alignment, and N is the number of elements that the matrix contained; R_i is the element of the first matrix, and S_i is the element of the second matrix; and R and S are the average values of R_i and S_i, respectively[7].

The following steps were used to calculate the r values of NSCLC-related proteins.

Collection of homologous sequences—A search of the NCBI protein database was performed to obtain the protein sequences required to build the evolutionary trees. Each of these sequences then served as a template in a search for homologous protein sequences in other species, which was subjected to the condition $P(N)$<1 ×10^{-5} and performed using the related protein database provided by BLAST.

Generation of the distance matrix—First, a protein present in human NSCLC and homologous proteins identified in other species were classified into the same group. Then, groups of proteins were arranged using the ClustalW software, in which the exported arrangement results were adjusted to ensure that each protein in the inner groups had a similar order. Lastly, the Bioedit software was used to calculate the corresponding protein distance matrix for each group of proteins. Each matrix represents a different evolutionary relationship between homologous protein sequences within the groups.

Calculation of the corresponding correlation coefficient—Equation (2) was used to calculate the corresponding r value between NSCLC proteins, which was then used to measure the strength of the PPI.

Results

The prediction of NSCLC protein-protein interactions

The eligible homologous sequences of NSCLC-related proteins were obtained by searching the related protein databases. Ninety-two NSCLC protein sequences in Homo sapiens consisted of homologous sequences with $P(N)$ < 1 × 10^{-5}. We obtained a total of 4 186 pairs of PPIs by computation (Table 1).

As shown in Table 1, 67.2% of NSCLC-related PPIs possessed an r value greater than 0.82, and 77.3% possessed an r value greater than 0.50, indicating that most of the proteins would interact strongly with each other.

The PPI network that was developed included 92 NSCLC-related proteins (Figure 1). Unfortunately, most of the proteins interacted closely, which made it difficult to divide them into functional modules. According to the Information Theory, information that is too complicated cannot provide valid or useful data.

Therefore, an alternate approach was adopted in which only the proteins with an r value less than 0.50 were analyzed. We identified the protein interactions one at a time. Fifteen specific proteins were found to possess predominantly similar interactions, including Cyclin D1 (CCND1), E-cadherin (CDH1), Cyclin-dependent kinase inhibitor 2A (CDKN2A), chemokine (C-X-C motif) ligand 12 (CXCL12), epidermal growth factor (EGF), epidermal growth factor receptor (EGFR), TNF receptor superfamily, member 6 (FAS), FK506

Table 1. Interaction coefficient (r) statistic on protein-protein interactions (PPIs) of non-small cell lung cancer (NSCLC)

r	−0.447	−0.289	−0.13	0.028	0.186	0.344	0.503	0.661	0.819	0.977	>0.977
Frequency	1	7	153	293	170	308	196	198	268	1348	1224
Rate	0.0002	0.0017	0.0367	0.0703	0.0408	0.0739	0.0470	0.0475	0.0643	0.3236	0.2938
Cumulative frequency	0.0002	0.0019	0.0386	0.1090	0.1498	0.2237	0.2708	0.3183	0.3826	0.7062	1.0000

binding protein 12-rapamycin associated protein 1 (FRAP1), O-6-methylguanine-DNA methyltransferase (MGMT), parkinson protein 2, E3 ubiquitin protein ligase (PARK2), phosphatase and tensin homolog (PTEN), calcium channel voltage-dependent alpha 2/delta subunit 2 (CACNA2D2), tubulin beta class I (TUBB), SWI/SNF-related, matrix-associated, actin-dependent regulator of chromatin, subfamily a, member 2 (SMARCA2), and wingless-type MMTV integration site family, member 7A (WNT7A). The functions and annotations of these proteins were retrieved (Table 2).

Although the majority of the 15 proteins possessed an *r* value less than 0.50, the analysis demonstrated that some calculated r values were less than 0.10 or

Table 2. Functions and annotations of 15 specific proteins of NSCLC

Protein	Function	Annotation
CCND1	Regulatory component of the Cyclin D1-CDK4 (DC) complex that phosphorylates and inhibits members of the retinoBLASToma (RB) protein family including RB1 and regulates the cell cycle during G_1/S transition.	G_1/S-specific Cyclin D1
CDH1	A calcium-dependent cell-cell adhesion glycoprotein comprised of five extracellular cadherin repeats, a transmembrane region, and a highly conserved cytoplasmic tail.	Adherin 1, type 1 preproprotein
CDKN2A	Capable of inducing cell cycle arrest in G_1 and G_2 phases, acts as a tumor suppressor.	Cyclin-dependent kinase inhibitor isoform 1
CXCL12	A very important factor in carcinogenesis and the neovascularization linked to tumor progression.	Chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)
EGF	A growth factor that plays an important role in the regulation of cell growth, proliferation, and differentiation.	Epidermal growth factor receptor isoform a
EGFR	A major regulatory protein in normal cellular processes such as proliferation, differentiation, and development.	Epidermal growth factor receptor isoform a
FAS	Binding with its receptor induces apoptosis. Fas ligand/receptor interactions play an important role in the regulation of the immune system and the progression of cancer.	FK506-binding protein 12-rapamycin associated protein 1
FRAP1	A serine/threonine protein kinase that regulates cell growth, cell proliferation, cell motility, cell survival, protein synthesis, and transcription.	FK506-binding protein 12-rapamycin associated protein 1
MGMT	Involved in the cellular defense against the biological effects of O^6-methylguanine (O^6-MeG) in DNA.	Methylated DNA-protein-cysteine methyltransferase
PARK2	A more general protein in the ubiquitin proteasomal pathway by participating in the removal of abnormally folded or damaged protein.	E3 ubiquitin-protein ligase parkin
SMARCA2	A member of the SWI/SNF family of proteins. Members of this family have helicase and ATPase activities and are thought to regulate transcription of certain genes by altering the chromatin structure around those genes. Two transcript variants encoding different isoforms have been found for this gene, which contains a trinucleotide repeat (CAG) length polymorphism.	Probable global transcription activator SNF2
WNT7A	A member of the WNT gene family, which consists of structurally related genes that encode secreted signaling proteins. These proteins have been implicated in oncogenesis and in several developmental processes, including regulation of cell fate and patterning during embryogenesis. Mutations in this gene are associated with Fuhrmann and Al-Awadi/Raas-Rothschild/Schinzel phocomelia syndromes.	Wingless-type MMTV integration site family, member precursor
PTEN	The protein negatively regulates intracellular levels of phosphatidylinositol-3,4,5-trisphosphate in cells and functions as a tumor suppressor by negatively regulating AKT/PKB signaling pathway.	Phosphatidylinositol-3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN
CACNA2D2	Local in the voltage-dependent calcium channel complex. Calcium channels mediate the influx of calcium ions into the cell upon membrane polarization and consist of a complex of alpha-1, alpha-2/delta, beta, and gamma subunits in a 1:1:1:1 ratio.	Voltage-dependent calcium channel subunit alpha-2/delta-2
TUBB	Making up microtubules, and has GTP enzymes, the dimers bound to GTP tend to assemble into microtubules, while dimers bound to GDP tend to fall apart; thus, this GTP cycle regulated by tubulin, beta class is essential for the dynamic instability of the microtubule.	Tubulin, beta class

The texts above were retrieved from the database of NCBI.

negative. Surprisingly, there were exceptions in which some proteins had larger r values when evaluated with other specific proteins. For example, the r values between PARK2 and CCND1, CDKN2A, EGFR, and FAS were all greater than 0.36, whereas most of the r values between PARK2 and other proteins were less than 0.1 or were negative. Only the r value between TUBB and CXCL12 was greater than 0.35. Although the r value between CACNA2D2 and CDH1 reached 0.5889, most of the r values between CACNA2D2 and other proteins were negative. Each of these exceptional cases (including proteins with an r value greater than 0.35) were present in the 15 proteins mentioned previously (Table 3). To provide experimental evidence that supports the PPI prediction, we researched the Biological General Repository for Interaction Datasets (BioGRID) and PubMed databases and found experiments that confirmed the PPIs between EGFR-EGF, PARK2-FAS, PTEN-FAS, and CACNA2D2-CDH1.

The results of the analysis indicated that these 15 proteins could be categorized into a special family, which we defined as F(K). The proteins of F(K) may possess unique associations that contribute to the development of NSCLC, and it is likely that these proteins are members of a distinct pathway (Table 2). We built a PPI sub-network for F(K) based on the data in Table 3.

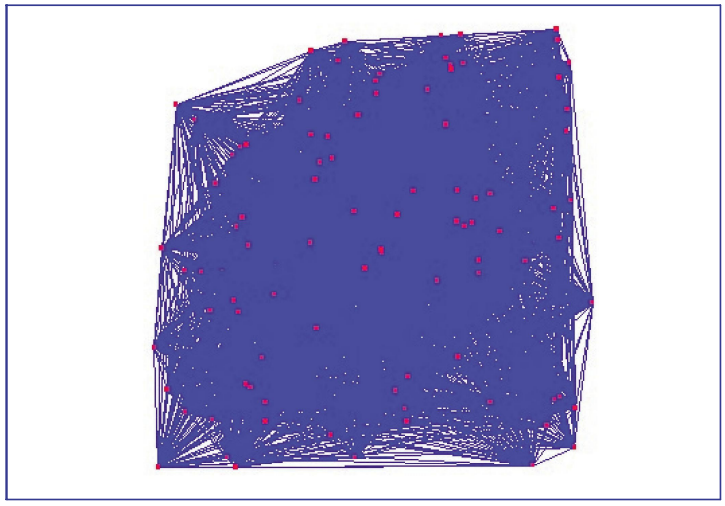

Figure 1. **Protein-protein interaction network of 92 non-small cell lung cancer–related proteins**

Table 3. Proteins with an unusually large *r* value within F(K) and some of the experimental confirmed PPIs by retrieving the BioGRID and Pubmed databases

Protein of F(K)	Interaction protein	*r*	Protein of F(K)	Interaction protein	*r*
CACNA2D2	CDH1*	0.5889	CDKN2A	CCND1	0.9136
EGFR	CCND1	0.9347	SMARCA2	CCND1	0.8923
	CDKN2A	0.8263		CDKN2A	0.9871
	EGF*	0.3766		EGF	0.7245
FRAP1	CCND1	0.9938		EGFR	0.8189
	CDKN2A	0.8661		FRAP1	0.8376
	EGFR	0.9261		MGMT	0.6020
PARK2	CCND1	0.3696	TUBB	CXCL12	0.3587
	CDKN2A	0.3865	WNT7A	CCND1	0.9127
	EGFR	0.3792		CDKN2A	0.9872
	FAS*	0.4110		EGFR	0.8554
PTEN	CDH1	0.8420		FRAP1	0.8635
	FAS*	0.7596		MGMT	0.5929
	PARK2	0.5224		SMARCA2	0.9932

The "*" marked EGFR-EGF, PARK2-FAS, PTEN-FAS and CACNA2D2-CDH1 had been validated by retrieving BioGRID and PubMed databases.

The PPI sub-network consisted of several key nodes, which included PARK2, WNT7A, SMARCA2, FRAP1, CDKN2A, CCND1, and EGFR (Figure 2). Each of the 7 proteins possessed more than 5 interactions. These 7 proteins may play important roles in the development of NSCLC, and these proteins may serve not only as valuable molecular markers for the early diagnosis of lung cancer but also as potential targets for the clinical treatment of lung cancer.

Inferring lung cancer development mode combined with the PPI sub-network

In 1998, through a prospective molecular pathology research, Sekine *et al.*[8] defined an occurrence process of genetic variation mode for lung cancer. Sekine believed that lung cancer first developed from the gene loss in the chromosome 3p region, followed by the genetic mutation of the 9p region (the P16 gene is located in the 9p12 region). Lastly, modification of the P53 or *RAS* gene occurs[8].

In the current study, the NCBI databases were thoroughly searched to identify the chromosomal localization of the genes encoding the proteins of interest (Figure 2). However, the genetic mutation mode in which lung cancer occurred and developed was found to be particularly complicated (Table 4). Upon evaluation of the 7 key node proteins (Figure 2), the occurrence and development of lung cancer was linked not only with gene loss in the 3p region (WNT7A, 3p25) and genetic mutation in the 9p region but also with similar events in the regions of 1p36.2 (FRAP1), 6q25.2-q27 (PARK2), and 11q13 (CCND1). Lastly, the invasion or metastasis of lung cancer happened (P15, P16, C-MYC, C-erbB-2, and EGFR genetic mutations, which include 10q and additional regions).

Discussion

Presently, the most challenging aspect of bioinformatics is the prediction of PPIs. However, compared with most experimental approaches, simulation and calculation approaches cost less and enable predictions to be made more efficiently. In recent years, many bioinformatics methods have been adapted to PPI calculations and have been categorized into four types of methods: methods based on genomic information; methods based on evolutionary relationships; methods

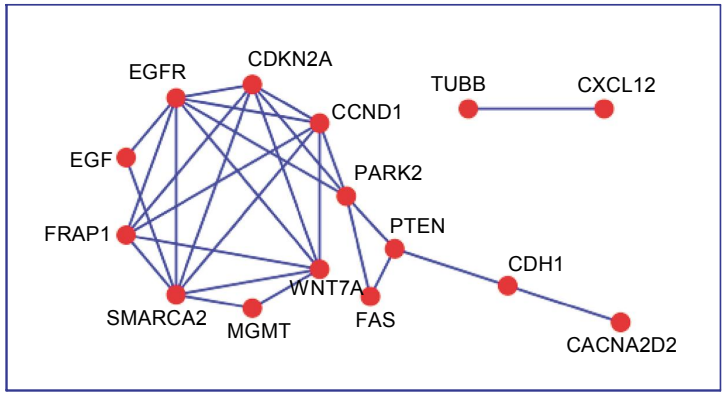

Figure 2. The protein-protein interaction sub-network of F(K)

Table 4. Genes chromosomal localization of F(K) encoding proteins

Chromosome	Genes and locations
1p	FRAP1 at 1p36.2
3p	CACNA2D at 3p21.3, WNT7A at 3p25
4q	EGF at 4q25
6q	TUBB at 6q21.33, PARK2 at 6q25.2–q27
7p	EGFR at 7p12
9p	CDKN2A at 9p21, SMARCA2 at 9p22.3
10q	CXCL12 at 10q11.1, PTEN at 10q23.3, FAS at 10q24.1, MGMT at 10q26
11q	CCND1 at 11q13
16q	CDH1 at 16q22.1

based on protein sequences (*ab initio*); and methods based on protein 3-dimensional structural information. However, the primary problems are inconsistencies that arise from comparing methods that use different databases and the use of databases that may be unreliable[9].

The systems biology approach has been applied to study the early molecular mechanisms of lung cancer to identify specific molecular markers with high sensitivity. These specific molecular markers are of great significance in molecular staging, early diagnosis, prediction of recurrence after tumor resection, correct judgment of prognosis, and choosing appropriate treatment programs for lung cancer patients.

Based on the Coevolution Theory, we predicted PPIs that occur in NSCLC and generated a total of 4 186 pairs of PPI data via performing calculations with 92 protein sequences. It was found that 77.3% of the r values of protein interaction associated with NSCLC were greater than 0.50, indicating that most of the proteins had strong interactions with each other. By adopting an alternate approach, 15 key proteins were identified and categorized into a special protein family, F(K), which included CCND1, CDH1, CDKN2A, CSCL12, EGF, EGFR, FAS, FRAP1, MGMT, PARK2, PTEN, CACNA2D2, TUBB, SMARCA2, and WNT7A.

The PPI sub-network of F(K) was constructed. A preliminary analysis determined 7 key node proteins: PARK2, WNT7A, SMARCA2, FRAP1, CDKN2A, CCND1, and EGFR.

The PPIs of EGFR-EGF, PARK2-FAS, PTEN-FAS, and CACNA2D2-CDH1 were confirmed by retrieving information from the BioGRID and PubMed databases[10]. EGFR-EGF participates in the regulation of receptor endocytosis[11] and receptor phosphorylation at Ser1039 and Thr1041. EGFR mutants exhibit enhanced tyrosine kinase activity in response to EGF and increased sensitivity to inhibition by gefitinib[12]. Additionally, PARK2-FAS regulates CDK4 activity via a novel CDK4-binding protein and appears to function as a growth factor sensor that may facilitate the formation and activation of Cyclin D-CDK complexes in the presence of inhibitory levels of INK4 proteins[13]. These proteins may play important roles in the development of NSCLC. These proteins may potentially serve as molecular markers for the early diagnosis and targets for the clinical treatment of lung cancer[14]. Accordingly, research has demonstrated that EGFR and FAS are very important molecular markers for NSCLC[15].

With reference to the development mode of lung cancer established by Sekine *et al*.[8], we proposed a detailed evolution mode that incorporated the results of the PPI sub-network. This mode indicates that lung cancer may arise due to the loss of *WNT7A* in the 3p25 region, followed by the genetic mutation of *CDKN2A* in the 9p21 region and *SMARCA2* in the 9p22.3 region. Lung cancer might also be linked to the loss of genes or genetic mutations that occur in the 1p36.2 region (*FRAP1*), the 6p25.2–q27 region (*PARK2*), or multiple mutations in the 10q region and the 11q13 region (*CCND1*). Subsequently, the genetic mutation of *P53* or *RAS* would then occur[16].

Currently, the coevolution approach is widely applied to predict PPIs. However, the analysis on the interaction prediction predominantly focused on prokaryotes. There are several key assumptions that the coevolution approach applies to predict the PPIs of eukaryotes. First, compared with the proteome of single-cell prokaryotes, the eukaryotic proteome consists of more proteins and more complex PPIs. Under these conditions, it becomes difficult to analyze and predict the whole interactome. Therefore, human NSCLC-related proteins were selected because they were fewer in number, making it relatively easier to analyze the protein sequences. The biological functions of these proteins predominantly include cell communication, signal transduction, and cell cycle regulation[17]. Understanding these interactions is very important for researching cancer detection and treatment. Second, the eukaryotic genome is very large, and the complete genomic sequences of vertebrate species are limited. Due to the limited sources of homologous protein sequences, it would be difficult to build multi-gene trees. Such restrictions would affect the study scope and accuracy of the coevolution approach used in the proteomic analysis of eukaryotes. Third, the phylogenetic relationships of eukaryotes, especially those of higher animals, would have greater similarity. There would exist only small differences between proteomes, and no gene would exhibit the phenomenon of horizontal transfer of non-normal genetic variation. These features make the application of the coevolution method more simple, more accurate and reliable.

This study demonstrates the need for more direct experimental evidence to validate the PPIs calculated by the coevolution method because the results may include a high rate of false positive or negative data. Additionally, the calculation methods need to be further improved.

Acknowledgments

This work was supported by 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: 2012-06-26;
accepted: 2012-06-30.

References

[1] Panagopoulos N, Apostolakis E, Koletsis E, et al. Low incidence of bronchopleural fistula after pneumonectomy for lung cancer. Interact Cardiovasc Thorac Surg, 2009,9:571–575.

[2] Wang J, Rao S, Chu J, et al. A protein interaction network for pluripotency of embryonic stem cells. Nature, 2006,444:364–368.

[3] Lim J, Hao T, Shaw C, et al. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell, 2006,125:801–814.

[4] Kar G, Gursoy A, Keskin O. Human cancer protein-protein interaction network: a structural perspective. PLoS Comp Bio, 2009, 5:e1000601.

[5] Rual JF, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature, 2005,437:1173–1178.

[6] Fares MA, McNally D. CAPS: coevolution analysis using protein sequences. Bioinformatics, 2006,22:2821–2822.

[7] Gobel U, Sander C, Schneider R, et al. Correlated mutations and residue contacts in proteins. Proteins, 1994,18:309–317.

[8] Sekine I, Minna JD, Nishio K, et al. Genes regulating the sensitivity of solid tumor cell lines to cytotoxic agents: a literature review. Jpn J Clin Oncol, 2007,37:329–336.

[9] Buck MJ, Atchley WR. Networks of coevolving sites in structural and functional domains of serpin proteins. Mol Biol Evol, 2005,22:1627–1634.

[10] Sucharita B, Sudha KP, Misako W, et al. FAS expression inversely correlates with PTEN level in prostate cancer and a PI 3-kinase inhibitor synergizes with FAS siRNA to induce apoptosis. Oncogene, 2005,24:5389–5395.

[11] Tatematsu A, Shimizu J, Murakami Y, et al. Epidermal growth factor receptor mutations in small cell lung cancer. Clin Cancer Res, 2008,14:6092–6096.

[12] Lynch TJ, Bell DW, Sordella R, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. New Engl J Med, 2004,350:2129–2139.

[13] Nicholson RI, Gee JM, Harper ME, et al. EGFR and cancer prognosis. Eur J Cancer, 2001,37 Suppl 4:S9–S15.

[14] Brusevold IJ, Aasrum M, Bryne M, et al. Migration induced by epidermal and hepatocyte growth factors in oral squamous carcinoma cells *in vitro*: role of MEK/ERK, p38 and PI-3 kinase/Akt. J Oral Path Med, 2012, doi: 0.1111/j.1600 – 0714.2012.01139.x

[15] Lee SH, Shin MS, Park WS, et al. Alterations of Fas (Apo-1/CD95) gene in non-small cell lung cancer. Oncogene, 1999, 18:3754–3760.

[16] Shapiro GI. Cyclin-dependent kinase pathways as targets for cancer treatment. J Clin Oncol, 2006,24:1770–1783.

[17] Brown MPS, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Nat Acad Sci U S A, 2000,97:262–267.